

**MICRO-EXPRESSION RECOGNITION ANALYSIS  
USING FACIAL STRAIN**

**LIONG SZE TENG**

**FACULTY OF COMPUTER SCIENCE  
AND INFORMATION TECHNOLOGY  
UNIVERSITY OF MALAYA  
KUALA LUMPUR**

**2017**

**MICRO-EXPRESSION RECOGNITION ANALYSIS  
USING FACIAL STRAIN**

**LIONG SZE TENG**

**THESIS SUBMITTED IN FULFILMENT  
OF THE REQUIREMENTS  
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**FACULTY OF COMPUTER SCIENCE  
AND INFORMATION TECHNOLOGY  
UNIVERSITY OF MALAYA  
KUALA LUMPUR**

**2017**

# UNIVERSITI MALAYA

## ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **Liong Sze Teng**

Registration/Matrix No.: **WHA140008**

Name of Degree: **Doctor of Philosophy**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

**Micro-expression Recognition Analysis using Facial Strain**

Field of Study: **Image Processing**

I do solemnly and sincerely declare that:

- (1) I am the sole author/writer of this Work;
- (2) This work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- (5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date

Subscribed and solemnly declared before,

Witness's Signature

Date

Name:

Designation:

## ABSTRACT

Facial micro-expression analysis has attracted much attention from the computer vision and psychology communities due to its viability in a broad range of applications, including medical diagnosis, police interrogation, national security, business negotiation, and social interactions. However, the micro and subtle occurrence that appears on the face poses a major challenge to the development of an efficient automated micro-expression recognition system. Therefore, to date, the annotation of the ground-truths (i.e., emotion label, onset, apex and offset frame indices) are still performed manually by psychologists or trained experts. This thesis briefly reviews the conventional automatic facial micro-expression recognition methods and their related works. In general, an automatic facial micro-expression recognition system consists of three basic steps, namely: image pre-processing, feature extraction, and emotion classification. This thesis mainly focuses on the enhancement of the first two steps over conventional methods in the literature. Specifically, a hybrid facial regions selection for pre-processing is proposed. This method is able to eliminate some parts of the face that are irrelevant to any facial emotions. Then, an effective feature descriptor, namely, optical strain, is utilized to capture the variations in characteristics and properties of the micro-expressions in the video. Next, a feature descriptor is developed to encode the essential expressiveness of the apex frame because the information of a single apex frame exhibits the highest variation of motion intensity, which is adequate to represent the emotion of the entire video. Finally, this thesis is concluded by highlighting its contributions and limitations, as well as suggesting possible future directions related to micro-expression recognition system.



## ABSTRAK

Analisis mikro-ekspresi pada wajah telah menarik banyak perhatian dari komuniti visi komputer dan psikologi, disebabkan oleh kegunaannya dalam pelbagai aplikasi termasuk diagnosis perubatan, soal siasat polis, keselamatan negara, perundingan perniagaan dan interaksi sosial. Namun, kejadian mikro-ekspresi yang kecil dan halus telah menjadi cabaran utama dalam usaha pembangunan system pengiktirafan mikro ekspresi automatik yang cekap. Setakat ini, anotasi “ground-truth” (iaitu label emosi, indeks bingkai permulaan, puncak dan pengakhiran) masih dibuat secara manual oleh ahli psikologi atau pakar yang terlatih. Disertasi ini mengkaji dengan ringkas kaedah-kaedah konvensional pengiktirafan mikro-ekspresi wajah automatik dan kerja-kerja yang berkaitan. Secara umumnya, sistem pengiktirafan mikro-ekspresi wajah automatik terdiri daripada tiga langkah, iaitu: pra-pemprosesan imej, pengekstrakan ciri dan klasifikasi emosi. Disertasi ini memberi tumpuan kepada penambahbaikan dua langkah pertama daripada kaedah konvensional dalam kesusasteraan. Terutamanya, satu teknik pemilihan kawasan wajah hibrid untuk pra-pemprosesan telah dicadangkan. Kaedah ini dapat menghapuskan bahagian wajah yang tidak berkaitan dengan emosi. Di samping itu, satu deskriptor ciri yang berkesan, iaitu ketegangan optik, digunakan untuk menangkap sifat-sifat dan ciri-ciri perubahan mikro-ekspresi dalam video. Selain itu, satu deskriptor ciri telah dicipta untuk mengekod ekspresi yang penting dalam bingkai puncak sahaja kerana maklumat bingkai puncak menunjukkan intensiti gerakan tertinggi dan ia tersesuai digunakan untuk mewakili emosi keseluruhan video. Akhir sekali, disertasi ini membuat kesimpulan tentang sumbangan dan had-had kajian ini, serta cadangan untuk hala tuju masa depan sistem pengiktirafan mikro-ekspresi.

## ACKNOWLEDGEMENTS

First of all, I would like to extend my most sincere gratitude to my supervisor Dr. KokSheik Wong and co-supervisor Assoc. Prof. Keat Keong Phang, Prof. Raphael Phan (Multimedia University) and Dr. John See (Multimedia University) for their invaluable guidance and assistance throughout the course of my Ph.D studies. They have always been encouraging, helpful and giving good advice whenever needed. This study would never have been completed Without their support and dedicated involvement in every research stage.

I would like to express my greatest appreciation to 2beAware group members for their support in this research. In particular, I am deeply grateful to Dr. Le Ngo Anh Cat, Dr. Yan Dan Wang and Yee Hui Oh. I wish to thank them for their patience in explaining the research ideas in detail and guiding me in developing the algorithms from scratch.

My sincere thanks also goes to my MSPIH group members for providing training and supervision in order to improve my presentation skills. They offered insightful comments, suggestions and discussions that undoubtedly helped to facilitate my research progress.

I am also indebted to my fellow colleagues at the Faculty of Engineering of MMU, including Dr. Vishnu Monn Baskaran, Dr. Ivan Ku, and Wei Zhe Lee for their company throughout this

I would like to thank Telekom Malaysia Research and Development (TM R&D) and University Malaya Research Collaboration Grant for financial support in the publication of conference and journal manuscripts.

Last but not the least, I owe my deepest thanks to my family members, specially my ever supporting parents, for all of the sacrifices that you have made on my behalf.

## TABLE OF CONTENTS

Original Literary Work Declaration.....	ii
Abstract.....	iii
Abstrak.....	iv
Acknowledgements.....	v
Table of Contents .....	vi
List of Figures .....	x
List of Tables.....	xiv
List of Symbols and Abbreviations.....	xvi
List of Appendices .....	xviii
 <b>CHAPTER 1: INTRODUCTION .....</b>	 <b>1</b>
1.1 Understanding Micro-expression.....	1
1.2 General framework of a Micro-expression Recognition System.....	3
1.3 Problem Statements .....	4
1.4 Objectives.....	5
1.5 Scopes and Limitations .....	6
1.6 Contributions .....	6
1.7 Structure of Thesis .....	8
 <b>CHAPTER 2: LITERATURE REVIEW .....</b>	 <b>9</b>
2.1 Overview .....	9
2.2 Overview of Image Pre-processing .....	9
2.2.1 Face Registration and Alignment .....	9
2.2.2 Image Filtering .....	11
2.2.3 Facial Region Selection.....	14
2.3 Overview of Feature Extraction.....	16
2.3.1 Local Binary Pattern.....	17

2.3.2	Local Binary Pattern on Three Orthogonal Planes.....	22
2.3.3	Optical Flow .....	27
2.3.4	Optical Strain.....	31
2.4	Overview of Micro-expression Databases .....	36
2.4.1	SMIC .....	36
2.4.2	SMIC II.....	38
2.4.3	CASME .....	47
2.4.4	CASME II.....	48
2.4.5	Other Micro-expression Databases .....	50
2.5	Summary and Limitations.....	52
<b>CHAPTER 3: HYBRID FACIAL REGIONS SELECTION.....</b>		<b>55</b>
3.1	Overview .....	55
3.2	Motivation .....	55
3.3	Literature Review .....	56
3.3.1	Action Unit.....	56
3.3.2	Region of Interest .....	61
3.3.3	Landmark Coordinate Detector .....	63
3.3.4	Feature Representation .....	65
3.3.5	Databases .....	65
3.4	Proposed Facial Regions Selection .....	66
3.4.1	Hybrid RoIs Extraction Approach .....	69
3.4.2	Optical Strain Features (OSF) feature extractor .....	71
3.4.3	Block-based LBP-TOP feature extractor .....	76
3.5	Experiments .....	78
3.5.1	Datasets .....	78
3.5.2	Experiment Settings .....	78
3.6	Results and Discussions.....	80
3.6.1	Parameter Analysis.....	80

3.6.2	Recognition Performance .....	83
3.6.3	Discussion on Computational Cost .....	87
3.7	Summary .....	90
<b>CHAPTER 4: FEATURE EXTRACTION BASED ON FACIAL STRAIN .....</b>		<b>91</b>
4.1	Overview .....	91
4.2	Literature Review .....	92
4.2.1	Optical Strain .....	92
4.2.2	Block-based LBP-TOP .....	92
4.2.3	Pooling .....	93
4.2.4	Image Filtering .....	93
4.3	Proposed Algorithm .....	94
4.3.1	Optical Strain Features .....	95
4.3.2	Optical Strain Weighted Features .....	100
4.3.3	Concatenating OSF and OSW Features .....	104
4.4	Experiments .....	105
4.4.1	Datasets .....	105
4.4.2	Setup .....	105
4.5	Results and Discussions .....	107
4.5.1	Detection and Recognition Results .....	107
4.5.2	Discussions .....	109
4.5.3	Comparison with Other Spatio-temporal Features.....	113
4.6	Summary .....	116
<b>CHAPTER 5: FEATURE EXTRACTION USING APEX FRAME .....</b>		<b>118</b>
5.1	Overview .....	118
5.2	Introduction .....	119
5.3	Motivation .....	120
5.4	Literature Review .....	122

5.4.1	Apex Spotting in Short Videos.....	122
5.4.2	Micro-expression Spotting in Long and Short Videos .....	122
5.4.3	Micro-expression Spotting and Recognition in Long Videos .....	123
5.4.4	Eye Blinking Issue in Long Videos .....	124
5.4.5	Feature Extraction and Face Representation .....	124
5.5	Proposed Algorithm .....	126
5.5.1	Apex Frame Spotting in Short Video .....	126
5.5.2	Apex Frame Spotting in Long Video .....	131
5.5.3	Micro-expression Recognition .....	134
5.6	Performance Metrics.....	139
5.7	Results and Discussions.....	140
5.7.1	Short Videos .....	140
5.7.2	Long Videos .....	147
5.8	Summary .....	152
5.9	Prima Facie .....	154
<b>CHAPTER 6: CONCLUSION.....</b>		<b>155</b>
6.1	Summary .....	155
6.2	Limitations.....	156
6.3	Future Works.....	157
Appendices.....		159
References.....		161

## LIST OF FIGURES

Figure 1.1: The top row of images show some examples of macro-expression (from CK+ database) and the second row of images show examples of micro-expression (from CASME II database). The emotion types are: (a-b) happiness, (c-d) surprise, (e-f) sadness, (g-h) disgust and (i-j): fear .....	2
Figure 1.2: Block diagram of a facial micro-expression recognition system .....	4
Figure 2.1: Face transformation. (a) Model face with feature points detected; (b) Sample face before transformation; (c) Results of mapping the feature points from sample face to model face.....	11
Figure 2.2: A face image before and after applying the filters. (a) Original image; (b) Gaussian filter; (c) Wiener filter; (d) Sobel filter .....	15
Figure 2.3: Sub-regions selected for the facial expressions: (a) Anger; (b) Disgust; (c) Fear; (d) Joy; (e) Sadness; (f) Surprise; (g) Neutral .....	16
Figure 2.4: Example of three different circularly symmetric neighbor sets, $[P,R]$ : (a) $[4,1]$ ; (b) $[8,1]$ ; (c) $[8,2]$ .....	18
Figure 2.5: Illustration of processes in the basic LBP operator .....	19
Figure 2.6: Block-based LBP: (a) Face image; (b) Face equally divided into $5 \times 5$ blocks; (c) Histogram for each block; (d) Resultant feature histogram .....	21
Figure 2.7: Neighborhood topology of LBP for bio-imaging application: (a) Circular; (b) Ellipse; (c) Parabola; (d) Hyperbola; (e) Archimedean spiral .....	22
Figure 2.8: Block-based LBP-TOP features extraction of the first two block volumes from a video sequence: (a) Block volumes; (b) LBP features from three orthogonal planes; (c) Histogram concatenation of each block volume from $XY$ , $XT$ and $YT$ planes to form a single histogram; (d) Histogram concatenation from the two block volumes ....	24
Figure 2.9: Optical flow estimation of the moving object between temporally-consecutive images towards the directions of: (a) Left; (b)Upper right.....	28
Figure 2.10: Formation of HOOF with four bins.....	30
Figure 2.11: Optical flow and optical strain computed between the onset and apex frames. Visualization of (a) Horizontal optical flow; (b) Vertical optical flow; (c) Optical strain .....	32

Figure 2.12: The eyes are masked for privacy concerns. Face is segmented into: (a) Three regions (i.e., forehead, cheeks, and mouth) (Shreve et al., 2009); (b) Eight regions (i.e., forehead, left and right of eye, left and right of cheek, left and right of mouth and chin) (Shreve, Godavarthy, et al., 2011); (c) Four regions (i.e., upper left, lower left, upper right and lower right) (Shreve et al., 2014) .....	35
Figure 2.13: (a) SMIC sample images; (b) Video mapping on the curve by adopting TIM to produce a new video .....	39
Figure 2.14: The acquisition setup for micro-expression elicitation of SMIC II database .....	40
Figure 2.15: Example of short and long videos with onset, apex and offset annotations.....	41
Figure 2.16: The acquisition setup for micro-expression elicitation of CASME II database.....	50
Figure 3.1: Example of AUs of the FACS and their interpretations, adapted from (Y. Zhang & Ji, 2005) .....	58
Figure 3.2: An example of a micro-expression video with onset-apex-offset frame annotation, showing a ‘Surprise’ expression.....	59
Figure 3.3: Examples of the emotion and AU labels (facial movements are highlighted) in CASME II database (Yan, Li, et al., 2014) : (a) Fear - AU 20; (b) Sadness - AU 1; (c) Disgust - AU L4; (d) Happiness - AU R12; (e) Surprise - AU R2 .....	61
Figure 3.4: Regions of interest suggested by: (a) Zhong et al. (Zhong et al., 2015); (b) Happy and Routray (Happy & Routray, 2015); (c) Anderson and McOwan (Anderson & McOwan, 2006); (d) Wang et al. (S. J. Wang, Yan, et al., 2014) .....	63
Figure 3.5: Example of annotating the 66 landmark coordinates using DRMF method on a CASME II image .....	64
Figure 3.6: Flowchart of the proposed hybrid regions of interest extraction method. ....	68
Figure 3.7: Cropping out the three RoIs: (a) The 66 landmark points marked by DRMF; (b) The rectangular boxes are set based on the coordinate of the 12 landmark points of the four borders .....	71
Figure 3.8: Optical strain feature extraction for the first video after cropping out the RoIs: (a) Original frames, $f_{1,j}$ ; (b) Strain maps, $m_{1,j}$ ; (c) Temporal pooled strain map; (d) Maximum-normalized and resized frame .....	74
Figure 3.9: Comparison of the normalized optical strain magnitude between the three RoIs (taken together) and the entire face region along a sample video sequence .....	75



Figure 3.10: Features of the first two blocks volumes extracted by using block-based LBP-TOP: (a) Each RoI is partitioned into $3 \times 3$ blocks; (b) LBP features generated from $XY, XT$ and $YT$ planes; (c) Concatenation of features in each block into a single histogram; (d) Concatenation of block histograms to form the final histogram.....	77
Figure 3.11: A sample video sequence of ‘Surprise’ micro-expression from SMIC-HS dataset.....	79
Figure 3.12: Results (Percentage of improvement in accuracy over baseline) of various combination of parameter settings by holding the value of: (a) $w$ in OSF, $r \in [15, 20]$ ; (b) $w$ in LBP-TOP, $N \in [3, 4]$ ; (c) $r$ in OSF, $w \in [6, 10]$ , and; (d) $N$ in LBP-TOP, $w \in [6, 10]$ .....	81
Figure 3.13: The percentage of improvement in recognition accuracy achieved by varying parameters $w$ with $r$ or $N$ using: (a) OSF method for SMIC-HS; (b) LBP-TOP method for SMIC-HS; (c) OSF method for CASME II; (d) LBP-TOP method for CASME II.....	86
Figure 4.1: Overview of the proposed algorithm.....	95
Figure 4.2: Effect of $\tau_l$ and $\tau_u$ values on micro-expression recognition rate for the SMIC-HS database .....	97
Figure 4.3: Example of vertical segmentation of the optical strain frame into three regions .....	98
Figure 4.4: Extracting OSF from a sample video sequence: (a) Original images, $f_{1,j}$ ; (b) Optical strain maps $m_{1,j}$ ; (c) Images after pre-processing; (d) Temporal pooled strain image; (e) Normalized and resized strain image .....	99
Figure 4.5: OSW histogram formation: (a) Each $j$ -th frame in $m_{1,j}$ is divided into $5 \times 5$ blocks before the values of $\epsilon_{x,y}$ within each block region are spatially pooled; (b) The block-wise strain magnitudes $z_{b_1,b_2}$ from all frames ( $j \in 1 \dots F_{i-1}$ ) are temporally mean pooled; (c) The weighting matrix $\mathbf{W}$ of size $N \times N$ is formed; (d) Coefficients of $\mathbf{W}$ are multiplied by their respective $XY$ -plane histogram bins .....	101
Figure 4.6: Top row: a sample image from SMIC-HS (left) and the corresponding optical strain map (right). Bottom row: a sample image from CASME II (left) and the corresponding optical strain map (right). Noise block at the bottom left and right corners are marked	103
Figure 4.7: Micro-averaged accuracy results of the baseline (LBP-TOP) and OSW methods using different LBP-TOP radii parameters on SMIC-HS database based on LOSOCV .....	112
Figure 4.8: Recognition accuracy results of the baseline (LBP-TOP) and OSF + OSW methods using different block partitions in LBP-TOP. The baseline results are denoted by the dashed lines (with transparent fill) ...	113

Figure 5.1: Flowchart of the apex frame spotting and emotion recognition system...	119
Figure 5.2: An example of a long and short video with annotated ground-truth labels indicating the onset, apex and offset frames .....	121
Figure 5.3: Flow diagram of apex frame spotting in short video.....	126
Figure 5.4: Demonstration of the apex frame spotting in a video sequence using LBP feature extractor with <i>divide &amp; conquer</i> strategy .....	130
Figure 5.5: Flow diagram of apex frame spotting in long video.....	131
Figure 5.6: Eye masking process: (a) There are 6 landmark coordinates which marked the boundaries of the left (landmark points 37, 38, 39, 40, 41 and 42) and 6 on the right (landmark points 43, 44, 45, 46, 47 and 48) eye regions; (b) The eye regions are removed after adding some pixel margins.....	132
Figure 5.7: Illustration of extraction of the three RoIs: (a) 66 landmark coordinates labeled by DRMF; (b) The four edges (i.e., top, bottom, left and right) are determined based on the landmark point locations; (c) Each RoI is partitioned into four blocks with the same size	133
Figure 5.8: Obtaining Bi-Weighted Oriented Optical Flow (Bi-WOOF) features.....	135
Figure 5.9: Bi-WOOF features formation: (a) $\theta$ and $\rho$ images are divided into $N \times N$ blocks. In each block, the values of $\rho$ for each pixel are treated as local weights to be multiplied with their respective $\theta$ histogram bins; (b) It forms a locally weighted HOOF with feature size of $N \times N \times C$ ; (c) $\zeta_{b1,b2}$ denotes the global weighting matrix, which is derived from $\varepsilon$ image; (d) Finally, $\zeta_{b1,b2}$ are multiplied with their corresponding locally weighted HOOF .....	138
Figure 5.10: Illustration of components derived from optical flow using the apex and first frames of a video: (a) Horizontal vector of optical flow, $p$ ; (b) Vertical vector of optical flow, $q$ ; (c) Orientation, $\theta$ ; (d) Magnitude, $\rho$ ; (e) Optical strain, $\varepsilon$ .....	145
Figure 5.11: Top row without eye masking, bottom row with eye masking: (a-b) First frame in the video; (c-d) Spotted apex frame; (e-f) Ground-truth apex frame; (g-h) Plots of optical strain magnitudes across the video sequence. Relevant frames are marked.....	151

## LIST OF TABLES

Table 2.1: Recognition accuracy in CASME II using different combination of radii values with fixed block size and neighboring points .....	27
Table 2.2: Detailed information of the SMIC database .....	38
Table 2.3: Detailed information of the SMIC-HS, SMIC-VIS and SMIC-HR datasets .....	42
Table 2.4: Detailed information of the SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR datasets .....	45
Table 2.5: Detailed information of CASME A and CASME B databases.....	48
Table 2.6: Detailed information of the CASME II and CASME II-RAW databases..	51
Table 2.7: General information of the USF-HD, Polikovsky's and YorkDDT databases .....	51
Table 3.1: Emotion description in terms of facial action units .....	69
Table 3.2: Frequency of the face regions based on the action units for five emotions	69
Table 3.3: The landmark points determining the corresponding RoIs bounding boxes .....	70
Table 3.4: Reproduced baseline recognition results (%).....	84
Table 3.5: F-measure, recall and precision scores of the proposed <i>RoI-selective</i> approach against their respective baselines in four different scenarios (averaging across the good parameter ranges in Figure 3.13)....	88
Table 3.6: Comparison of recognition results of the proposed method to existing methods in measurements of Accuracy, F-measure, recall and precision scores in LOSOCV protocol.....	89
Table 4.1: Micro-expression detection and recognition results on SMIC-HS and CASME II database with LBP-TOP of $5 \times 5$ block partitioning .....	107
Table 4.2: Micro-expression detection and recognition results on SMIC-HS and CASME II database with LBP-TOP of $8 \times 8$ block partitioning .....	107
Table 4.3: Confusion matrices of baseline and OSF + OSW methods for detection task on SMIC-HS database with LBP-TOP of $5 \times 5$ block partitioning .....	108
Table 4.4: Confusion matrices of baseline and OSF + OSW methods for recognition task on SMIC-HS database with LBP-TOP of $8 \times 8$ block partitioning .....	109

Table 4.5: Confusion matrices of baseline and OSF + OSW methods for recognition task on CASME II database with LBP-TOP of $5 \times 5$ block partitioning .....	109
Table 4.6: F-measure, recall and precision scores for detection and recognition performance on SMIC-HS and CASME II database with LBP-TOP of $5 \times 5$ block partitioning.....	110
Table 4.7: F-measure, recall and precision scores for detection and recognition performance on SMIC-HS and CASME II database with LBP-TOP of $8 \times 8$ block partitioning.....	111
Table 4.8: Comparison of micro-expression detection and recognition accuracy results on the SMIC-HS and CASME II databases for different feature extraction methods .....	115
Table 5.1: Comparison of micro-expression recognition performance in terms of F-measure on the CASME II, SMIC-HS, SMIC-VIS and SMIC-NIR databases for the state-of-the-art feature extraction methods, random frame selection approach, and the proposed algorithm. ....	142
Table 5.2: Confusion matrices of baseline and Bi-WOOF (apex & first frame) for recognition task on the CASME II database .....	143
Table 5.3: Confusion matrices of baseline and Bi-WOOF (apex & first frame) for recognition task on the SMIC-HS database .....	144
Table 5.4: Performance of apex frame spotting with and without eye masking on the CASME II-RAW database measured by MAE. ....	148
Table 5.5: Performance of apex frame spotting with and without eye masking on the long videos databases measured by ASR.....	148
Table 5.6: Recognition performance for long videos in terms of F-measure .....	149
Table 5.7: Comparison of the recognition accuracy between the state-of-the-art method and the proposed method on the SMIC-E-VIS database .....	149
Table 5.8: Average number of frames in the short and long videos of the CASME II and three SMIC databases .....	152
Table 5.9: Confusion matrices for the recognition task on the CASME-II-RAW and SMIC-E-HS databases using the proposed method .....	153

## LIST OF SYMBOLS AND ABBREVIATIONS

$F_i$	: Total number of frames in a video.
$M$	: LBP-TOP Histogram.
$N \times N$	: Block partitioning.
$P$	: Number of neighbor points.
$R$	: Radii.
$X$	: Width of an image.
$Y$	: Height of an image.
$\bar{M}$	: Normalized LBP-TOP Histogram.
$\epsilon$	: Optical strain magnitude.
$\vec{u}$	: Displacement vector.
$f_{i,j}$	: Frame indices.
$m_{i,j}$	: Strain map.
$n$	: Total number of video sequence in the database.
$p$	: Horizontal component of optical flow vector.
$q$	: Vertical component of optical flow vector.
$s_i$	: Number of video in the database.
AAM	: Active Appearance Model.
ASM	: Active Shape Model.
ASR	: Apex Spotting Rate.
AU	: Action Unit.
Bi-WOOF	: Bi-Weighted Oriented Optical Flow.
CASME II	: Chinese Academy of Sciences Micro-Expression II.
CK+	: Extended Cohn Kanade.
CLM	: Constrained Local Model.
DRMF	: Discriminative Response Map Fitting.
DTSA	: Discriminant Tensor Subspace Analysis.
ELM	: Extreme Learning Machine.
ERI	: Expression Ratio Image.
FACS	: Facial Action Coding System.
FERET	: Facial Recognition Technology.
HOG	: Histogram of Gradient.
HOOF	: Histogram of Oriented Optical Flow.
LBP	: Local Binary Pattern.
LBP-SIP	: Local Binary Patterns with Six Intersection Points.
LBP-TOP	: Local Binary Pattern on Three Orthogonal Planes.
LOSOCV	: Leave-One-Subject-Out Cross Validation.

LOVOCV	: Leave-One-Video-Out Cross Validation.
LSTD	: Local Spatio-temporal Directional Features.
LWM	: Local Weighted Mean.
MAE	: Mean Absolute Error.
MKL	: Multiple Kernel Learning.
OF	: Optical Flow.
ORL	: Olivetti Research Laboratory.
OS	: Optical Strain.
OSF + OSW	: Concatenation the OSF and OSW.
OSW	: Optical Strain Weighted Features.
RF	: Random Forest.
riLBP	: Rotation Invariant LBP.
riuLBP	: Uniform Rotation Invariant LBP.
RoI	: Region of Interest.
SMIC	: Spontaneous Micro-expression.
SVM	: Support Vector Machine.
TIM	: Temporal Interpolation Model.
uLBP	: Uniform LBP.

## LIST OF APPENDICES

Appendix A: List of Publications and Papers Presented .....	159
---	-----

University of Malaya

## CHAPTER 1: INTRODUCTION

### 1.1 Understanding Micro-expression

Facial expression is the most common form of non-verbal communication that displays a person's feeling. It is also known as the universal language of emotion as it is shared among different cultures. There are six basic classes of emotions, notably happiness, surprise, anger, sad, fear and disgust (Ekman & Friesen, 1971). In general, facial expression can be categorized into two main types: macro-expression and micro-expression. Macro-expression, also known as the normal expression, is usually obvious and can be easily identified in real-time with the naked eye. However, since macro-expression is a voluntary expression, it can be exploited to deceive others by imitating or acting out the falsified expressions and portray them on the face. More precisely, macro-expression goes on and off on the face, normally between three quarters of a second to two seconds (Shreve et al., 2009), and can be found at multiple large areas of the face. Automatic macro-expression recognition is a popular research field and has been intensively studied over the past two decades.

On the other hand, in recent years, analysis of micro-expression has also attracted more and more attention in the field of computer vision. Yet, not many papers have been published. This is because micro-expression has shorter duration (micro) and lower intensity (subtle) (Ekman & Friesen, 1969). It often occurs at high speed and usually sustains within one-fifth to one-twenty-fifth of a second (Porter & Ten Brinke, 2008). Due to its extremely brief and rapid facial muscle movement, as well as the fact that it may only appear in a few small parts of the face, it is technically challenging to realize and recognize the micro-expressions in real-time conversations, except for the keen and trained observers (Ekman, 2009b).





**Figure 1.1:** The top row of images show some examples of macro-expression (from CK+ database) and the second row of images show examples of micro-expression (from CASME II database). The emotion types are: (a-b) happiness, (c-d) surprise, (e-f) sadness, (g-h) disgust and (i-j): fear

Micro-expressions are provoked involuntary and spontaneously. In other words, it is uncontrollable, and thus being able to reveal a person's concealed genuine feelings (Ekman & Friesen, 1969). Figure 1.1 illustrates some images containing either the macro- and micro-expressions. The attributes between macro- and micro-expressions can clearly be differentiated using the figure shown. Specifically, the characteristics of micro-expression makes it advantageous to be recognized and studied, as we are able to interpret whether someone is concealing his feeling or is trying to tell a lie (Ekman, 2009a). For instance, the suspects interrogated by the police can be caught lying. Analyzing a person's emotion can also help improve relationships and understand each other better. In addition, we can become more aware of our own emotions and manage them more effectively. In short, recognition of micro-expressions is beneficial in both our mundane lives and also society at large. It can be utilized in a wide range of applications, such as medical diagnosis, community safety, business negotiation, social interactions, etc. (Frank, Herbasz, et al., 2009; O'Sullivan et al., 2009; Frank, Maccario, & Govindaraju, 2009).

Micro-expressions have been studied intensively in the field of psychology. The first discovery of micro-expression was made by Haggard and Isaacs (Haggard & Isaacs, 1966), who named it as "micromomentary expression (MME)" and considered it as re-

pressed emotion. They stumbled upon it while performing analysis on a psychotherapeutic interview film, where they viewed the film frame-by-frame to understand the patient's thinking. They observed that the emotion shown by the patient changed dramatically within three to five frames. Few years later, Ekman and Friesen (Ekman & Friesen, 1969) introduced the term *micro-expression* from a case where the patient was trying to conceal his sad feeling by concealing it with a smile. The therapist detected his genuine feeling by carefully observing the subtle movements on his face. In fact, the patient was planning to commit suicide. In addition, they discovered that people who are trying to deceive others are more likely to attempt in disguising their facial behavior than body movements (Ekman & Friesen, 1974).

## **1.2 General framework of a Micro-expression Recognition System**

There are three main components in a general facial micro-expression recognition system, as illustrated in Figure 1.2. The input of the recognition system is the raw images, which are extracted from video sequences. The first stage of the system is pre-processing, which aims to enhance the image features for further analysis. Pre-processing includes smoothening the distorted image caused by illumination and lighting effects, normalizing the pixel intensity or scaling the brightness level distribution of the image affected by non-uniform light, face registration and alignment to transform and standardize all faces into a uniform size and shape based on a template face, etc. Next, the feature extraction process combines the large set of raw image data into a compact and discriminative feature vector. This process forms a compact representation of the image data by reducing the feature dimension. The resultant feature vector is able to describe the color, shape, texture and motion which exist in the image. It is also robust to translation, noise, occlusion, rotation, illuminations and scaling. The last step of the system is classification, which categorizes the emotion classes of the testing data based on the defined training



**Figure 1.2:** Block diagram of a facial micro-expression recognition system

data. It involves a learning algorithm that analyzes and organizes the training data into a finite desired group of clusters with respect to the emotion classes. Then, the emotion type of the testing data is determined based on the distance or similarity measure. This thesis focuses on the first two stages, namely pre-processing and feature extraction. In this research, several efficient and improved pre-processing and feature extraction techniques are proposed and implemented. Results show that they are capable of producing better recognition performances compared to the current state-of-the-art techniques.

### 1.3 Problem Statements

As previously explained, the study of micro-expression is still considered as a relatively new topic compared to that of macro-expression. For macro-expression, there are numerous well-established databases that are publicly available for benchmarking and evaluation purpose. Hence, many macro-expression recognition mechanisms have been designed to detect and classify a person's emotional state. Some of them are robust to operate in the wild (or the real world environment) and real-time scenario, while others manage to achieve perfect classification accuracies (around 100%) in certain databases.

On the contrary, the subtlety and minuteness of micro-expressions have profoundly hindered the progress of its related research works. To date, there are only a few micro-expression databases with proper data elicitation which provide complete ground-truths, due to the challenges of triggering the involuntary spontaneous expressions. This brings to an even greater obstacle in the development of micro-expression detection and recognition algorithms. In the literature, most state-of-the-art techniques achieve poor recognition accuracies of below 60%, even when they are tested on the entire short (or cropped)

videos which do not consist of micro-expression-unrelated motion. Furthermore, the pre-processing and feature extraction techniques employed in these systems are initially meant for macro-expression analysis, and may not perform well for micro-expressions due to different attributes and characteristics.

Until now, the labeling of ground-truths for micro-expressions are still performed manually by trained psychologists or professional experts. The drawbacks of the hand-labeled practice include: (a) inconsistent reliability - the labels may differ from person to person; (b) time consuming - requires great amount of effort and concentration for frame-by-frame evaluation; (c) costly - requires specific trained experts. The aforementioned issues indicate that it is essential to label the ground-truths automatically. Although there are some automatic micro-expression detection and recognition systems, there are still rooms for improvement. In summary, the problem statements for this research are as follows:

1. Poor performances by existing micro-expression detection and recognition systems due to the challenges in capturing the quick and minute micro-expressions.
2. Limited scope of exploration in pre-processing stage for micro-expression analysis.
3. Similar facial motion patterns appearing in consecutive frames implies redundancy and leads to a less discriminative set of features.

#### **1.4 Objectives**

This thesis aims to improve the current micro-expression detection and recognition systems by studying the pre-processing and feature extraction approaches. The primary objectives of this research are set out as follows:

1. To design a spatio-temporal feature extractor that can effectively describe the local appearance of micro-expressions.

2. To develop a pre-processing method that removes unwanted facial areas for concentrating only on regions that contribute meaningful micro-expression details.
3. To demonstrate that it is sufficient to consider only one frame out of the whole video sequence when extracting facial micro-expression features for recognition task.

### **1.5 Scopes and Limitations**

The scope and limitation of this study are as follows:

1. The types of emotion for micro-expression classification are strictly limited to only the emotions provided in the databases. No other types of emotion or subjects' feeling (i.e., anxiety, guilty, relief, etc.) are analyzed.
2. The databases used in the experiments are elicited under constrained laboratory conditions, which means that all the images have been pre-processed with face registration and alignment. There is no micro-expression database recorded in the wild.
3. The proposed pre-processing and feature extraction techniques are only tested on micro-expression videos.
4. To recognize the emotion type, only the muscle or skin tissue movements on the face are studied. Other features that may provide clues for micro-expressions, such as eye gaze and speech signals, are not considered.

### **1.6 Contributions**

Most existing pre-processing and feature extraction approaches applied on the micro expression videos are designed for macro-expression analysis. In this thesis, the weaknesses of these approaches are addressed and solutions are provided to mitigate the issues. The main contributions of this research are summarized as follows.

**RoI-Selective:** A hybrid facial region selection technique, **RoI-Selective** is proposed to extract several important parts of the face that contain significant and valuable micro-expression information. More precisely, it combines heuristic-based and automatic approaches to exclusively determine the salient facial regions. The heuristic-based approach relies statistically on the occurrence frequency of the facial action units for all the expressions, whereas the automatic detection of the landmark points is performed using a robust landmark detector. This work has been submitted to the Journal of Signal Processing Systems (JSPS).

**OSW + OSF:** For micro-expression detection and recognition, three distinct spatio-temporal feature extraction techniques are developed by utilizing optical strain magnitude. Optical strain is the extension of optical flow that has a higher order derivative, and possesses the ability to eliminate noises and preserve relative large facial muscle changes. For the three feature extraction techniques: (a) **OSW** - the optical strain magnitude is formulated as a weighting scheme to scale the values of the feature obtained by a baseline feature extractor called LBP-TOP; (b) **OSF** - the temporal details for each frame, which are derived from optical strain technique, are summed up to create a compact and efficient feature vector, and; (c) **OSW + OSF** - the two sets of features (i.e., **OSW** and **OSF**) are combined to form a more representative feature histogram. This work has been published in the Asian Conference on Computer Vision Workshops (ACCVW 2014), Intelligent Signal Processing and Communication Systems (ISPACS 2014) and Signal Processing: Image Communication (SPIC).

**Single Apex:** Apex frame in a micro expression video sequence is the frame that depicts the most expressive emotional state. Apex frame lies between the onset and offset frames, which are the instant of the beginning and ending of the micro-expression, respectively.

In this research, a single apex frame is utilized to represent the entire video. An automatic apex frame spotting approach is established to allocate the apex frame for various scenarios, including short and long videos, with and without onset and offset frame information. Moreover, a novel feature descriptor called **Bi-WOOF** is developed to encode the features obtained from the spotted apex frame. This work has been published in the IAPR Asian Conference on Pattern Recognition (ACPR2015). The extended work has been submitted to the Asian Conference on Computer Vision Workshop (ACCVW 2016) and Neurocomputing Journal.

## 1.7 Structure of Thesis

This thesis is organized into six chapters. The contents for each chapter, except for Chapter 1, are outlined briefly in this section. Chapter 2 provides literature review on existing image pre-processing and feature extraction algorithms, as well as micro-expression databases. In addition, the pros and cons for each method and database are discussed. In Chapter 3, a novel pre-processing method is proposed, which emphasizes on important facial regions to boost the micro-expression recognition performance. Chapter 4 describes the three distinct feature extractors, which have been devised from the optical strain technique. They are capable of encoding the subtle and short elapsed occurrence of micro-expressions. Apart from the conventional feature extractors that consider the entire video sequence or a sub-sequence of it for representation, Chapter 5 proposes a new approach, which expresses the features of a video based on only a single frame. Chapter 6 draws some conclusions and suggests possible future research directions.

## **CHAPTER 2: LITERATURE REVIEW**

### **2.1 Overview**

The materials in the literature are studied and presented in this chapter. Firstly, the techniques used in pre-processing are reviewed. Next, some of the popular feature extractors are discussed. Then, all the publicly available micro-expression databases are reviewed. Finally, the problems faced by the surveyed literature are identified and discussed.

### **2.2 Overview of Image Pre-processing**

As discussed in Chapter 1.2, it is essential to pre-process the raw micro-expression videos before conducting any experiment on the datasets. There are several pre-processing techniques in the literature which can effectively extract the image features resulting in better performances. Three widely-used pre-processing methods are studied, namely: (a) face registration and alignment; (b) image filtering, and; (c) facial region extraction. Each pre-processing methodology mentioned are detailed in the following sub-chapters.

#### **2.2.1 Face Registration and Alignment**

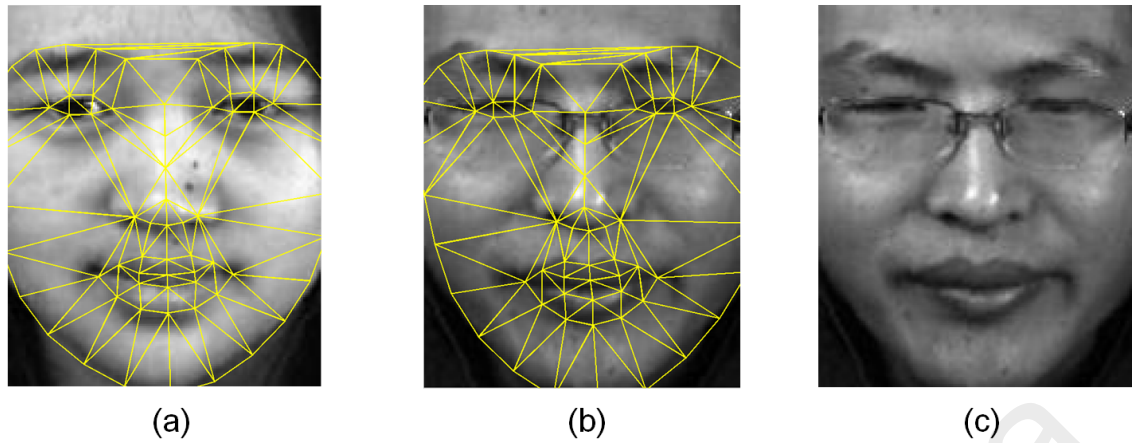
Due to the spontaneity of micro-expressions, the databases (i.e., Spontaneous Micro-expression (SMIC) (Pfister et al., 2011) and Chinese Academy of Sciences Micro-Expression II (CASME II) (Yan, Li, et al., 2014)) have gone through face registration and alignment processes before being released to the public. Although the datasets are collected under constrained laboratory condition, the external factors that contribute to feature noises, such as head movement and the shape of the face, need to be properly addressed.

There are four main steps involved in this process, namely:

1. Model face selection - a frontal face image with neutral expression is detected and selected as the model face (i.e., as the reference).



2. Landmark coordinate detection - a detector to allocate the facial landmark coordinates is adopted. Note that the terms *facial landmark coordinate* and *facial feature point* are used interchangeably in this thesis and will be selected based on the suitability of the context. Typically, there are three techniques commonly used in locating the facial feature points, namely: Active Shape Model (ASM) (Van Ginneken et al., 2002), Constrained Local Model (CLM) (Cristinacce & Cootes, 2006) and Active Appearance Model (AAM) (Cootes et al., 1998). ASM is employed to detect the facial landmark points in the micro-expression databases (i.e., SMIC and CASME II). It is a statistical model of the shape of an object that are repetitively deformed to fit a sample object. ASM is capable of spotting a total of sixty eight landmark coordinates on the face. The search commences from a mean shape aligned to the position and size of the face determined by a face detector. Then, the processes are iterated until convergence.
3. Face transformation - the first frame of each micro-expression sequence is normalized to the model face using a Local Weighted Mean (LWM) (Goshtasby, 1988) transformation. The rest of the frames are transformed and normalized using the same transformation matrix. The reason of using the same transformation matrix is that the occurrence of the micro-expression is too rapid, and the rigid head movement within the duration can be ignored. Additionally, the ASM landmark detector may not be sufficiently accurate thus it could cause a significant deviation of locations for the same points even when the face remains stationary. Figure 2.1 shows an example of face transformation.
4. Face normalization - the eye coordinates of the first frame for each normalized micro-expressions video clip are located and the face of each frame is cropped based on the rectangle determined by the eye coordinates.



**Figure 2.1:** Face transformation. (a) Model face with feature points detected; (b) Sample face before transformation; (c) Results of mapping the feature points from sample face to model face

### 2.2.2 Image Filtering

In addition to the noise introduced by the capturing device and environment, other factors including clothing and headset wire may affect the performance of a micro-expression recognition system. Since the motions characterized by the subtle facial expressions are very fine, it is likely that the presence of the unwarranted noises will generate false information to the micro-expression recognition system. Thus, a feasible pre-processing technique is required to suppress the unwanted noise to better describe the micro-expressions. Image filtering is a simple mathematical processing tool that is accomplished through a convolution operation. It is also a standard process that applies to almost all the image processing systems. Among the filtering processes, smoothing and blurring are the common ones, with adjustable levels to accommodate for different conditions and applications. Various types of filter are proposed in the literature, including Gaussian, Wiener, Sobel, Canny, Prewitt and Roberts. Three prominent image filters are chosen and discussed in this section, namely, Gaussian, Wiener and Sobel filters.

Gaussian filter is an adaptive filter controlled by a set of parameters based on an optimization algorithm (Forsyth & Ponce, 2002). It is a low pass filter as it removes high-frequency components from the image. The Gaussian filter formula can be expressed as

follows:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \quad (2.1)$$

where  $\sigma$  is the standard deviation of the Gaussian distribution. It is effective for removing Gaussian noises from image. In the work by Liu et al. (Z. Liu et al., 2001), Gaussian filter is adopted to minimize the noises on image in order to compute the change of illumination of a person or Expression Ratio Image (ERI) resulting from deformation of the person's face. The main reason Gaussian filter is chosen is that the degree of smoothing on the face can be tailored, where little smoothing is applied on expressional areas, while significant smoothing in the remaining areas. In other words, different treatment is applied to different region. Besides that, Gaussian filter is applied as the fundamental step for the cross-modality 2D-3D face recognition analysis (Jin et al., 2014). It is the pre-processing step for data enhancement by removing the noise spikes (i.e., the pixels that have exceptionally low or high pixel intensity values) on the face image collected. Despite the wide application of Gaussian filter, its smoothing process may reduce the fine image details. In this study of micro-expression recognition system, the subtle changes are particularly important. Thus, adjusting the Gaussian window to an optimal value is essential to efficiently filter out the noises while keeping the meaningful information about fine expression.

On the other hand, Wiener filter is a classic filter that is based on linear time-invariant estimation of an image (Goldstein et al., 1998). It is a low pass filter that finds the best reconstruction of a noisy signal by suppressing the overall mean square error in the inverse filtering and noise smoothing stages. It is able to remove noises that corrupt the signals. Wiener filter is often applied in the frequency domain. The image after taking the product

of the Wiener filter  $W(u, v)$  is estimated as follows:

$$\hat{G}(u, v) = W(u, v)O(u, v), \quad (2.2)$$

where  $O(u, v)$  is the Discrete Fourier Transform of the original image  $O(x, y)$ . In the application of digital image binarization to preserve useful texture information for low quality digitized documents, Wiener filter has demonstrated its efficiency in removing the noise areas by highlighting the contrast between background noise and text areas, while maintaining the desired image features (Gatos et al., 2004). In addition, Wiener filter has extended its efficacy to patch-based Wiener filter that can achieve near-optimal denoising (Chatterjee & Milanfar, 2012). It uses both the geometrically and photometrically similar patches to accurately estimate and learn the parameters. The denoising approach achieved promising performance on both grayscale and color images. It is worth noting that the parameters of the method are obtained from analytical formulation and they do not require further fine-tuning.

Last but not least, Sobel filter, also known as the Sobel-Feldman operator, has been widely used long before the inception of Gaussian filter and its derivatives. It is a pair of  $3 \times 3$  convolution masks that estimates the gradients of pixel intensity in the horizontal and vertical directions. Each pixel in the image is convolved by differentiating two rows or two columns to calculate the gradient. The Sobel gradient approximation for each pixel in an image is expressed as:

$$|G| = \sqrt{G_x^2 + G_y^2}, \quad (2.3)$$

where  $G_x$  and  $G_y$  are the approximated derivatives of the pixels in the horizontal and vertical directions. The main advantage of Sobel filter is its high sensitivity towards noises, enabling it to highlight the important edges in an image. Pai et al. (Pai & Chang, 2011)

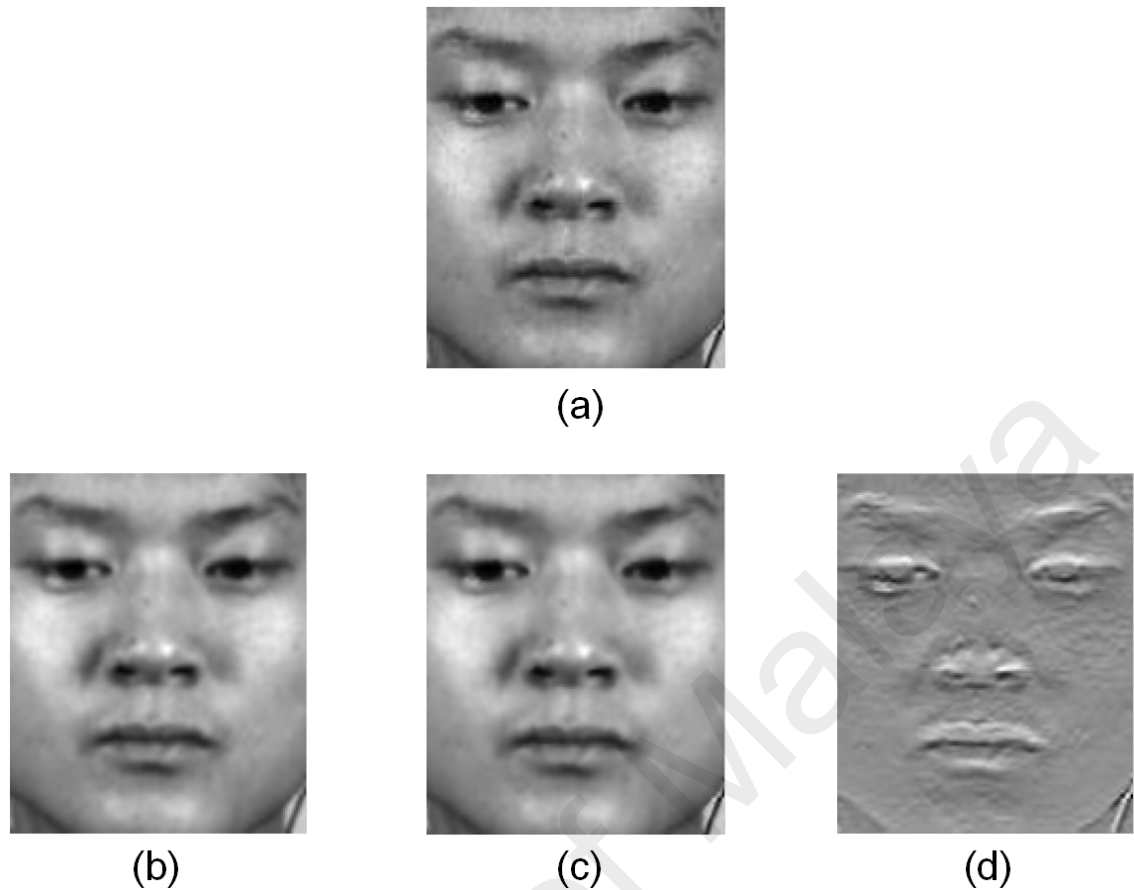
applied Sobel operator to produce a binary edge image (i.e., 1 indicates edge pixel and 0 denotes no edge) that has clear contrast between the face contour and background. The face and the background are successfully outlined accurately using the skin tone-segmenting processes. Furthermore, the dominant facial features such as the mouth and eyes regions are highlighted precisely, hence improving the subsequent recognition task. Another example of the application of Sobel filter is in skin color detection (Lamsal & Matsumoto, 2015). The skin color detector is modified by combining it with the Sobel edge detector (Viola & Jones, 2004). This method produces promising face detection performance, and is able to detect both the color and grayscale images in three facial image databases. However, Sobel filter is susceptible to noise. As the noise level increases, the magnitude of the edges decreases, thus affecting the edge detection accuracy.

Figure 2.2 illustrates the output of pre-processing the same image with three different filters (i.e., Gaussian, Wiener and Sobel filter) discussed above.

### **2.2.3 Facial Region Selection**

Many research papers demonstrated that extracting features from partial face is able to achieve better facial expression recognition performance, compared to when considering the entire face (Shan & Gritti, 2008; Fan & Verma, 2009). The main reason is that, these local facial patches contribute more expressional information towards distinguishing the expressions, and at the same time, eliminating the parts that do not correspond to the desired facial movements.

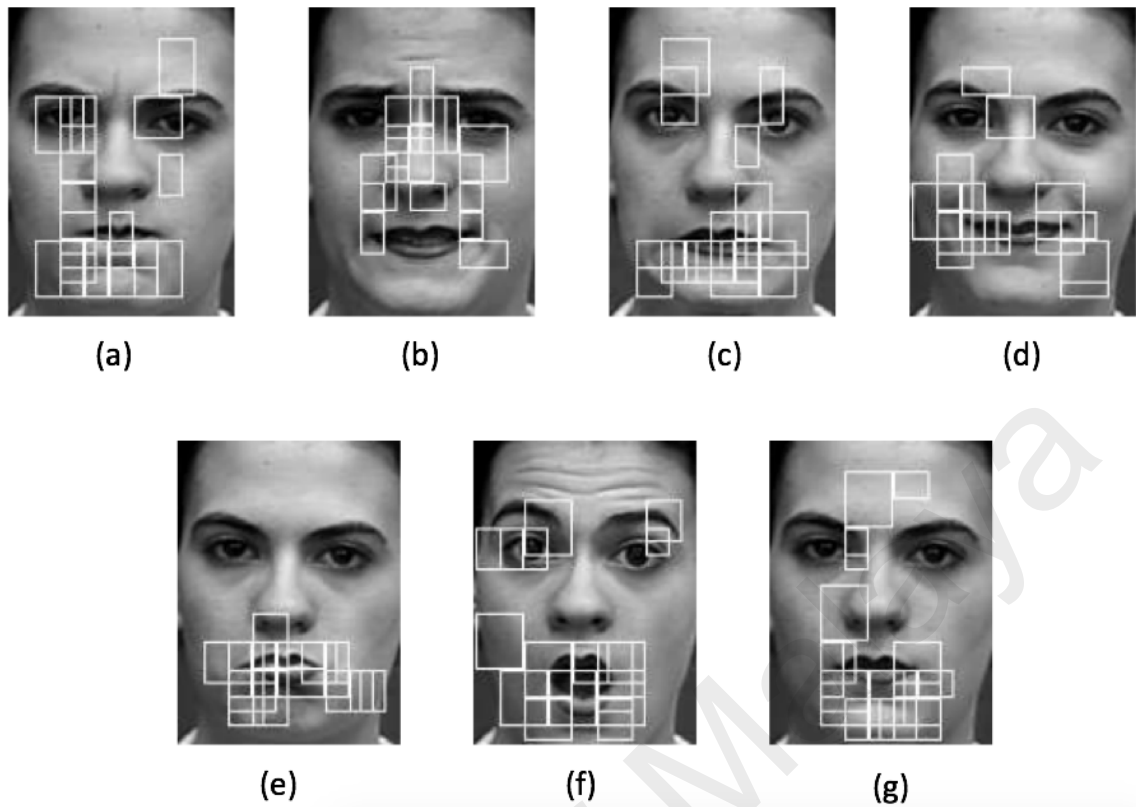
Shan and Gritti (Shan & Gritti, 2008) elicited a finite set of discriminative Regions of Interest (RoIs) that is more representative, instead of considering the entire face image. For their methodology, the face image is first divided into several sub-regions, before using the Local Binary Pattern (LBP) (Ojala et al., 1996) feature descriptor to capture the local appearance in region level. Finally, the weak classifier Adaboost (Schapire & Singer,



**Figure 2.2:** A face image before and after applying the filters. (a) Original image; (b) Gaussian filter; (c) Wiener filter; (d) Sobel filter

1999) is employed to learn and identify the most discriminative sub-regions (in term of LBP histogram). The method is tested on a facial expression database (i.e., Extended Cohn Kanade (CK+) (Lucey et al., 2010)), and attains consistently good recognition performances when three different classifiers are adopted. The paper summarized that the important features are mostly attributed to the eyes and mouth regions. Figure 2.3 shows the top 20 discriminative sub-regions selected for each individual expression. Nonetheless, the selection of the feature length is determined statistically based on the plot of recognition rate against the number of features.

Fan and Verma (Fan & Verma, 2009) discovered a combination of RoIs that contain the most important subtle facial motion information. These RoIs are “left eye region”, “right eye region”, “nose region” and “mouth region”. For analysis, the facial expression recognition performances are reported separately for each facial region. They discover



**Figure 2.3:** Sub-regions selected for the facial expressions: (a) Anger; (b) Disgust; (c) Fear; (d) Joy; (e) Sadness; (f) Surprise; (g) Neutral

that eyes and mouth achieve better recognition performance than the other facial regions, such as nose. By locating the most significant areas on the face, an excellent classification accuracy of 94% is obtained on the face benchmark database, i.e., Facial Recognition Technology (FERET) (Phillips et al., 1998). However, the size of the facial patches is sensitive to the feature representation and is completely based on empirically determined findings from numerous experiments.

### 2.3 Overview of Feature Extraction

Feature extraction process plays a vital role in a wide range of computer vision tasks, such as object recognition, image restoration, scene reconstruction, etc. Its main function is to reduce the image attributes from high dimensional feature space to a lower dimension. The resultant features possess more meaningful and richer information to better represent the original image, thereby facilitating the analysis of interest. In this sub-chapter, four popular low level feature descriptors, which are considered in this study, are described

in detail, namely: (a) LBP; (b) Local Binary Pattern on Three Orthogonal Planes (LBP-TOP); (c) Optical Flow (OF), and; (d) Optical Strain (OS). A low level feature refers to the component which directly deals with pixel intensities, for instance, color, texture, shape, edge and corner.

### 2.3.1 Local Binary Pattern

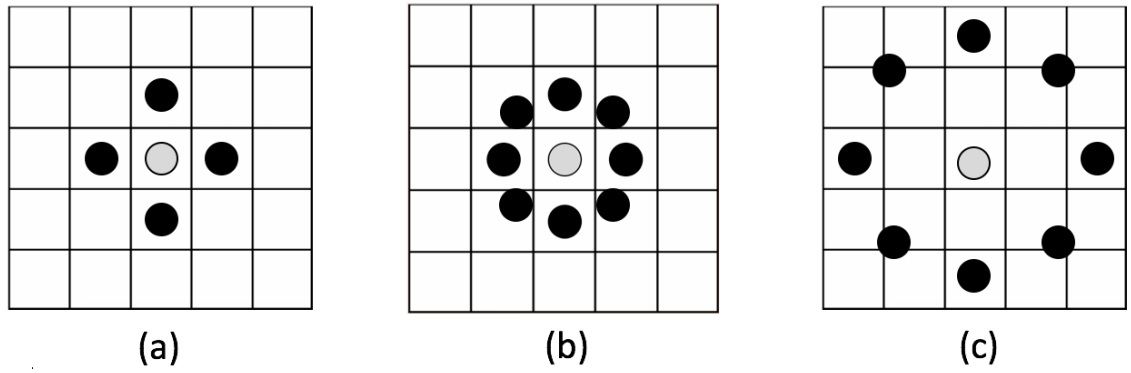
LBP feature descriptor, which was first introduced by Ojala et al. in 1996 (Ojala et al., 1996), is originally designed for texture analysis. It is an effective image feature descriptor, not only used for texture interpretation, but has also been further extended to various fields in computer vision due to its advantage of: (a) discrimination ability; (b) compact texture representation; (c) low computational complexity, as well as; (d) invariant to any monotonic gray-level changes. Examples of its application include face recognition (Xi et al., 2016), facial expression classification (Zhao & Pietikainen, 2007), event detection (Ma & Cisar, 2009) and even medical image analysis (Nanni et al., 2010; Mirmohamadsadeghi & Drygajlo, 2011).

LBP operator is a robust descriptor that represents a two-dimensional gray-scale image by converting pixel values into binarized pattern. The concept of LBP operator is simple, where the intensity value of the center pixel is compared to its circular neighboring pixels using a thresholding technique. Specifically, given a pixel  $c$  at position  $(x_c, y_c)$ , the binary code is computed by comparing the value of pixel  $c$  with its neighboring pixels:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p, s(x) = \begin{cases} 1, & x \geq 0; \\ 0, & x < 0, \end{cases} \quad (2.4)$$

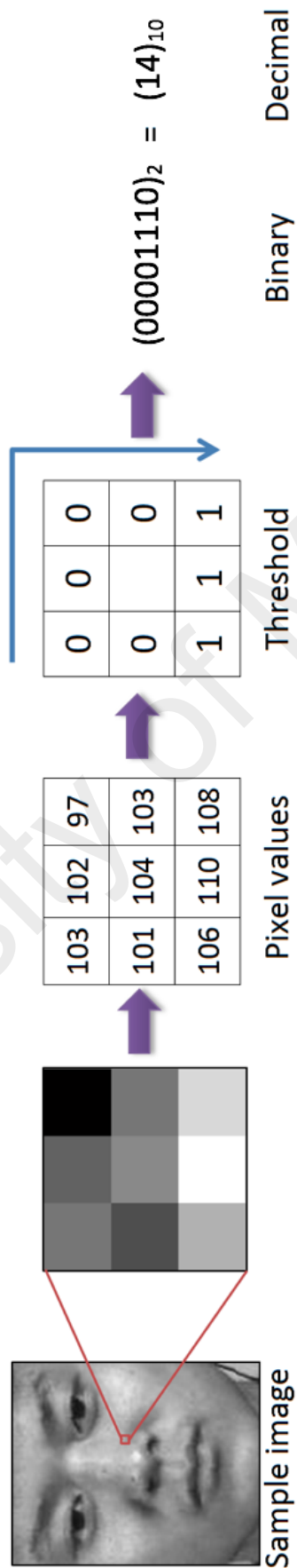
where  $P$  denotes the number of neighboring points equally sampled on a circle of radius  $R$ .  $g_c$  is the gray value of the center pixel and  $g_p$  is the gray value of the pixels equally sampled points around the circular neighborhood of radius  $R$ , and  $2^P$  is the weight cor-





**Figure 2.4:** Example of three different circularly symmetric neighbor sets,  $[P, R]$ : (a)  $[4, 1]$ ; (b)  $[8, 1]$ ; (c)  $[8, 2]$

responding to the neighboring pixels. Figure 2.4 shows different  $[P, R]$  sets of circular neighborhood. Figure 2.5 illustrates the basic LBP operator, where each pixel in the  $3 \times 3$  circular neighborhood is compared to pixel  $c$ . The  $2^P$  bins histogram is formed by calculating the occurrences of the LBP code derived across the whole face image.

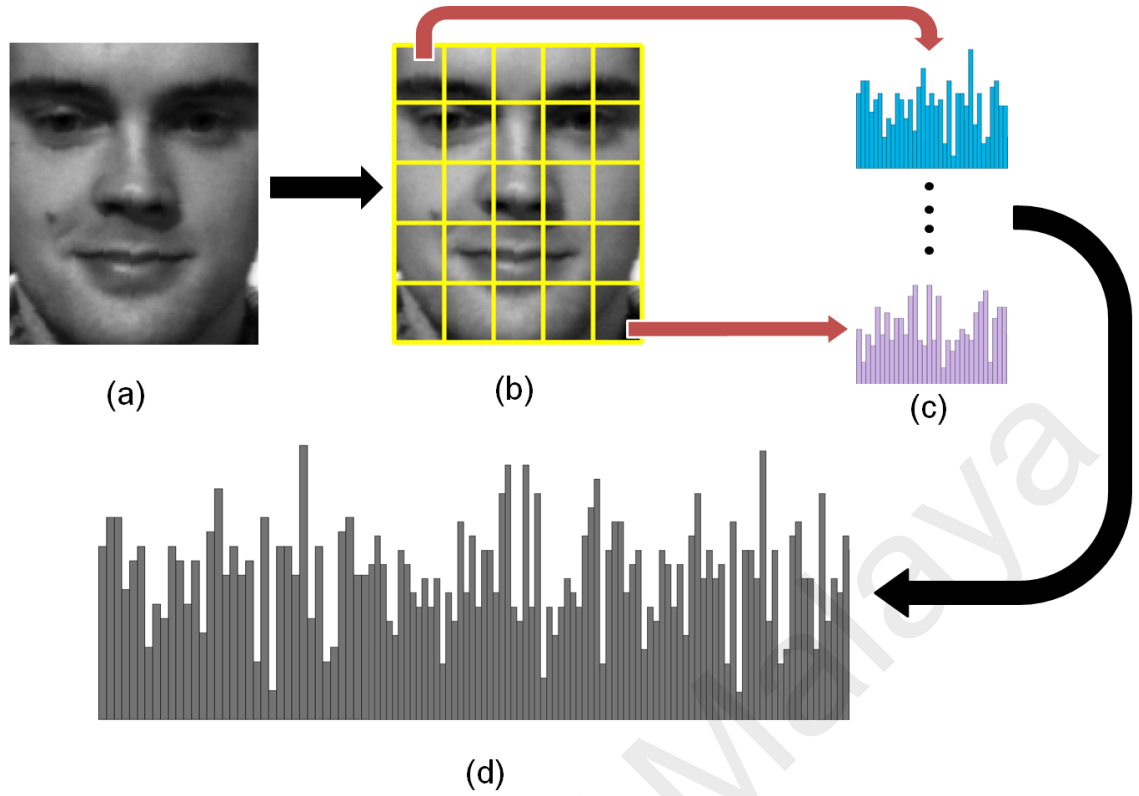


**Figure 2.5:** Illustration of processes in the basic LBP operator

LBP also provides different options for texture representation to reduce the number of histogram bins, such as *Uniform LBP (uLBP)*, *Rotation Invariant LBP (riLBP)* and *Uniform Rotation Invariant LBP (riuLBP)* (Ojala et al., 2002). Specifically, *uLBP* is a uniformity measure of pattern: if the total number of transitions from 0 to 1 or from 1 to 0 in the binary codes are less than or equal to 2, they are categorized as important. Otherwise, they are grouped into the miscellaneous bin in the histogram. For example, the patterns  $00000000_2$  (0 transition),  $10000000_2$  (1 transition) and  $01100000_2$  (2 transitions) are uniform whereas the patterns  $10001001_2$  (4 transitions) and  $11010010_2$  (5 transitions) are not. For *riLBP*, it is achieved using rotation invariant mapping, where the binary code circularly rotates until reaching the minimum value. For instance, the patterns  $010110000_2$  ( $176_{10}$ ),  $000101100_2$  ( $44_{10}$ ) and  $00001011_2$  ( $11_{10}$ ) are all mapped to the minimum code  $00001011_2$  ( $11_{10}$ ).

The block-based LBP was introduced in 2004. It equally divides the face area into small regions (Ahonen et al., 2004). This is to ensure that the spatial appearance can be described locally on a regional level. Then, the regional features are concatenated to form a global description of the face. Figure 2.6 shows the steps of constructing the block-based LBP feature histogram. The effectiveness of this approach has been validated on two face databases (i.e., FERET (Phillips et al., 1998) and Olivetti Research Laboratory (ORL) (Samaria & Harter, 1994)). When comparing to other approaches developed for face recognition, block-based LBP shows its superiority and robustness against different facial expressions, lighting condition and aging of the subjects.

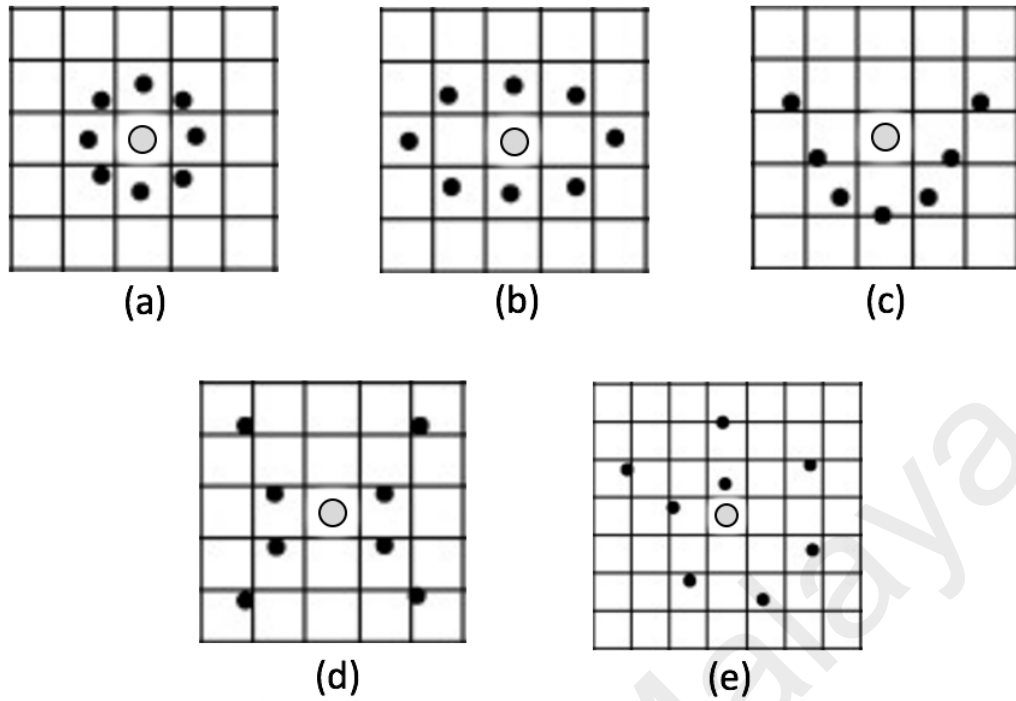
Nanni et al. (Nanni et al., 2010) presented several LBP variants by designing new encoding schemes for the validation of the local gray-scale differences. Instead of adopting the original circular shape when thresholding the center pixel to the surrounding pixels, four other neighborhood shapes are created, including, ellipse, parabola, hyperbola and archimedean spiral. The neighborhood topologies developed are shown in Figure 2.7.



**Figure 2.6:** Block-based LBP: (a) Face image; (b) Face equally divided into  $5 \times 5$  blocks; (c) Histogram for each block; (d) Resultant feature histogram

Their methods are tested on three medical databases: (a) Infant COPE database (Brahnam et al., 2007) - to categorize the pain states from neonatal images; (b) 2D-HeLa database (Chebira et al., 2007) - to classify the protein localization starting from fluorescence microscope images, and; (c) Pap smear database (Jantzen et al., 2005) - to detect abnormal smear cells. The elliptic neighborhood generates a reliable set of features in all three databases. Particularly in 2d-HeLa database, it outperformed all the texture descriptors in the literature.

The first work that applies LBP methodology on human detection task is demonstrated by Mu et al. (Mu et al., 2008). However, the ordinary LBP operator does not perform well in detecting the human in a personal album (i.e., INRIA human database (Dalal & Triggs, 2005)) due to high complexity and semantic consistency issues. Two feature descriptors, namely, Semantic-LBP and Fourier-LBP, are developed to address these problems. They are able to encode the features using a more compact representa-



**Figure 2.7:** Neighborhood topology of LBP for bio-imaging application: (a) Circular; (b) Ellipse; (c) Parabola; (d) Hyperbola; (e) Archimedean spiral

tion. Extensive experiments have been carried out to compare the methods with gradient-based feature descriptors. The better detection accuracy and higher computational speed achieved further confirm the feasibility of the binary codes in capturing meaningful local structures on image manifold.

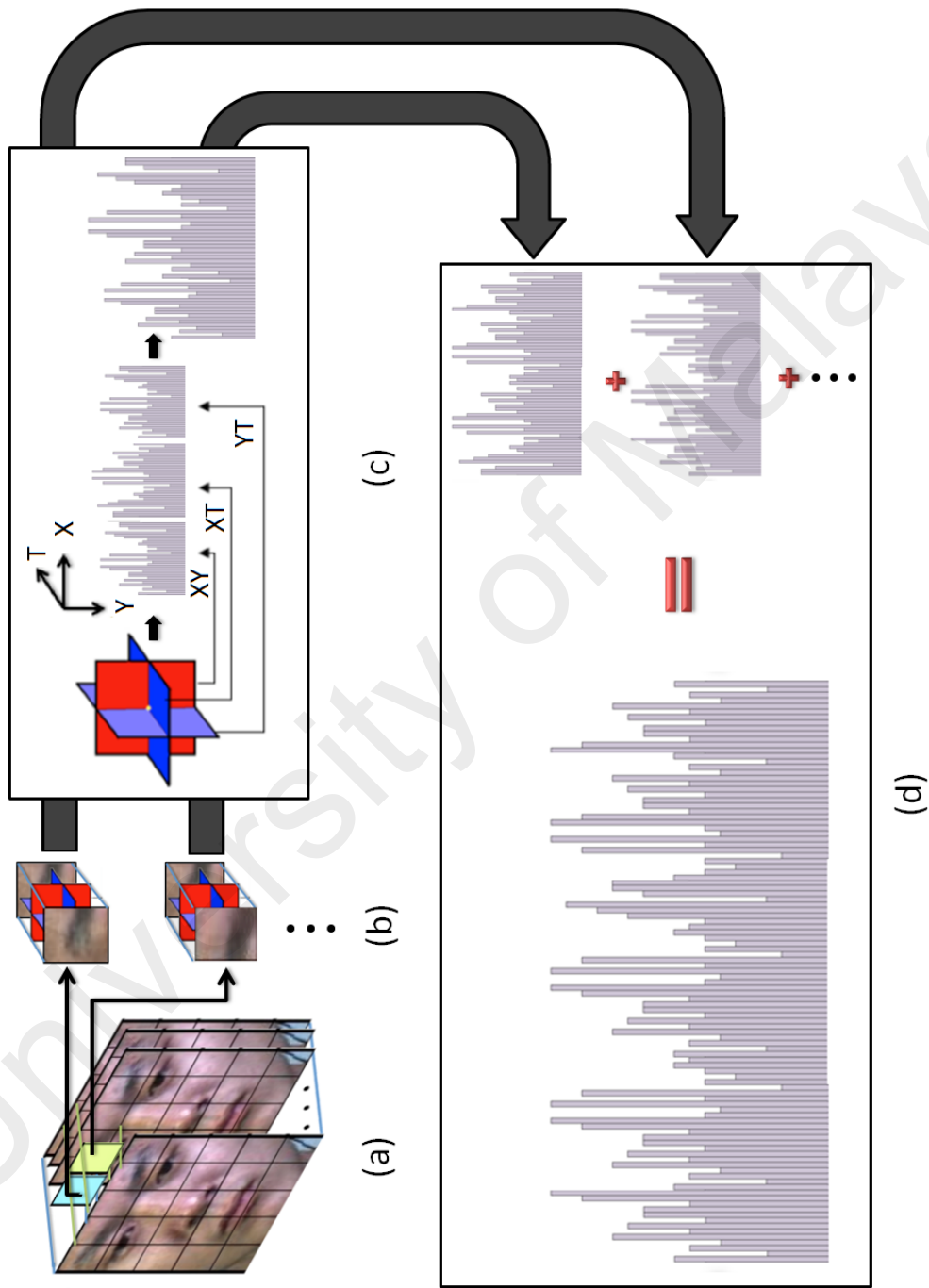
### 2.3.2 Local Binary Pattern on Three Orthogonal Planes

In 2007, Zhao and Pietikainen (Zhao & Pietikainen, 2007) proposed an extension of the spatial LBP to spatio-temporal volumes, namely LBP-TOP. The original LBP descriptor describes two dimensional local spatial structure of an image, whereas LBP-TOP can describe the three dimensional dynamic textures of a video sequence. Rather than extracting the 2D (i.e., from  $XY$  plane) features, the 3D variant of LBP considers the co-occurrence statistic on three orthogonal planes (i.e.,  $XY$ ,  $XT$  and  $YT$ ). The authors (Zhao & Pietikainen, 2007) listed several main advantages of LBP-TOP, namely: (a) capable of extracting the appearance and motion features in three directions; (b) ability to capture the local spatio-temporal transition information (i.e., pixel, region and volume levels); (c)

robustness to rotation, translation, illumination or skin color variation, gray-scale changes and even error in face alignment, as well as; (d) computational simplicity.

Similar to LBP, it has been extended to block-based LBP-TOP to describe more local attributes. Figure 2.8 shows the process of extracting the block-based LBP features from three orthogonal planes, followed by concatenating the computed local features into a resultant histogram.

University of Malaya



**Figure 2.8:** Block-based LBP-TOP features extraction of the first two block volumes from a video sequence: (a) Block volumes; (b) LBP features from three orthogonal planes; (c) Histogram concatenation of each block volume from  $XY$ ,  $XT$  and  $YT$  planes to form a single histogram; (d) Histogram concatenation from the two block volumes

The concatenated histogram describes the global motion of the face over a video sequence, and it can be succinctly denoted as  $M$ :

$$M_{b_1, b_2, d, c} = \sum_{x, y, t} I\{h_d(x, y, t) = c\} \quad (2.5)$$

where  $c \in \{0, \dots, 2^P - 1\}$ ,  $d \in \{0, 1, 2\}$ ,  $b_1, b_2 \in \{1, \dots, N\}$ , and  $2^P$  is the number of different labels produced by the LBP operator on the  $d$ -th plane for case  $d = 0$  refers to the  $XY$  plane, which describes the appearance; case  $d = 1$  refers to the  $XT$  plane, which describes the horizontal motion, and; case  $d = 2$  refers to the  $YT$  plane, which describes the vertical motion. As the LBP code cannot be computed at the borders of the 3D video volume, only the central part is taken into consideration.  $h_d(x, y, t)$  is the LBP code, i.e., Equation (2.4), of the central pixel  $(x, y, t)$  on the  $d$ -th plane, for  $x \in \{0, \dots, X - 1\}$ ,  $y \in \{0, \dots, Y - 1\}$ , and  $t \in \{0, \dots, T - 1\}$ .  $X$  and  $Y$  are the width and height of image (thus,  $b_1$  and  $b_2$  are the row and column indices, respectively), while  $T$  is the video length. As such, the functional term  $I\{A\}$  determines the count of the  $c$ -th histogram bin when  $h_d(x, y, t) = c$ :

$$I\{A\} = \begin{cases} 1, & \text{if } A \text{ is true;} \\ 0, & \text{otherwise.} \end{cases} \quad (2.6)$$

Hence, the final feature histogram is of dimension  $2^P \times 3N^2$ . For an appropriate comparison of the features among video samples of different spatial and temporal lengths, the concatenated histogram is sum-normalized to obtain a coherent descriptor  $\bar{M}$ :

$$\bar{M}_{b_1, b_2, d, c} = \frac{M_{b_1, b_2, d, c}}{\sum_{k=0}^{n_d-1} M_{b_1, b_2, d, k}}. \quad (2.7)$$



In this thesis, the LBP-TOP parameters are denoted by  $\text{LBP-TOP}_{P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T}$  where the  $P$  parameters indicate the number of neighbor points for each of the three orthogonal planes, while the  $R$  parameters denote the radii along the  $X$ ,  $Y$ , and  $T$  dimensions of the descriptor.

Unlike other feature extractors, LBP-TOP extracts the dynamic textures over space-time dimensions. Hence, it remains a popular choice of feature extraction method for various applications, including gait and action recognition (Kellokumpu et al., 2009; Mattivi & Shao, 2009), face spoofing (Chingovska et al., 2013), depression analysis (Joshi et al., 2012) and facial expression recognition (Zhao & Pietikäinen, 2009).

Furthermore, LBP-TOP has become a baseline feature extractor in micro-expression analysis due to its high discriminating power and impressive texture representation, for instance, in the databases: (a) SMIC (Pfister et al., 2011); (b) SMIC II (Li et al., 2013); (c) CASME (Yan et al., 2013); (d) CASME II (Yan, Li, et al., 2014). Specifically, LBP-TOP is utilized to perform detection and recognition tasks, where the former aims to classify micro-expressions versus non-micro-expressions, while the latter is to classify the emotions into different categories (i.e., happiness, surprise, disgust, sadness and others). Different combination of the parameters (i.e., block size, radii and number of neighboring points) in LBP-TOP setting generates different detection and classification performance. Yan et al. (Yan, Li, et al., 2014) provides the recognition performances of different radii values (i.e.,  $R_X$ ,  $R_Y$  and  $R_T$ ) with fixed  $5 \times 5$  block size and neighboring points of 4 in CASME II database, as tabulated in Table 2.1. Among all the values tested, the best performance obtained is 63.41% when the radii are  $R_X = 1$ ,  $R_Y = 1$  and  $R_T = 4$ .

Wang et al. (Y. Wang et al., 2014) extended the LBP-TOP feature descriptor by proposing the Local Binary Patterns with Six Intersection Points (LBP-SIP), before evaluating it in two micro-expression databases. LBP-SIP reduces redundant intersection points to encode less unwanted information, and hence leads to a more compact represen-

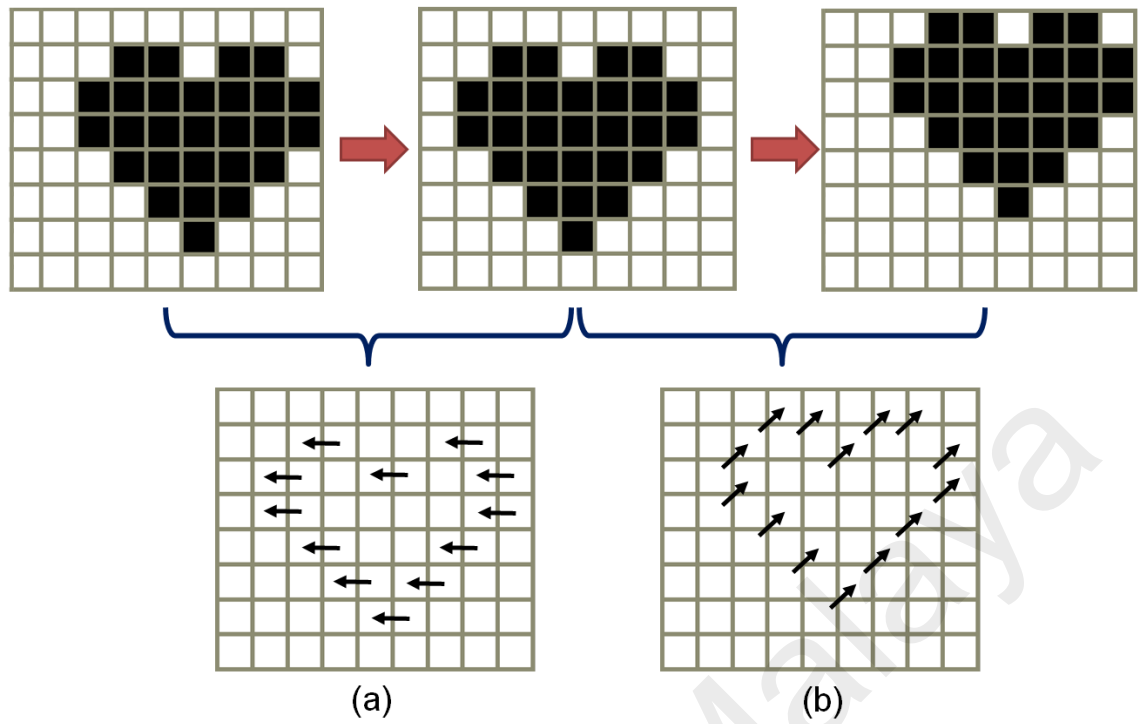
**Table 2.1:** Recognition accuracy in CASME II using different combination of radii values with fixed block size and neighboring points

$R_X$	$R_Y$	$R_T$	Recognition Accuracy (%)
1	1	2	63.01
1	1	3	62.60
<b>1</b>	<b>1</b>	<b>4</b>	<b>63.41</b>
2	2	2	61.38
2	2	3	61.79
2	2	4	62.20
3	3	2	58.54
3	3	3	61.79
3	3	4	58.55
4	4	2	58.94
4	4	3	61.38
4	4	4	60.57

tation and lower computational complexity. With a well-formed representation, LBP-SIP outperforms the original LBP-TOP in the SMIC II database. In addition, a Gaussian multi-resolution pyramid is incorporated to the LBP-SIP to capture important facial expressions in different resolutions. The resultant feature histogram that combines the face textures from each pyramid level produces a more superior recognition accuracy. However, the optimal value of the pyramid level that provides the best performance is determined empirically through experiments.

### 2.3.3 Optical Flow

Optical flow is one of the basic motion estimation techniques exploited in many computer vision applications to examine the motion pattern of object. The idea of optical flow was introduced in 1950 by a psychologist named Gibson (Gibson, 1950). Optical flow was first proposed to describe the visual perception and stimulus experienced by an animal when it moves through the world. Later on, optical flow stimulus has been expanded for the perception by the human observers during their own movement through the environment. It is the distribution of two-dimensional apparent motion of objects in an image based on the brightness patterns in an image (Horn & Schunck, 1981). It indicates the



**Figure 2.9:** Optical flow estimation of the moving object between temporally-consecutive images towards the directions of: (a) Left; (b)Upper right

displacement and velocity of each image pixel between temporally-consecutive images by assuming that the pixel intensity values are translated between subsequent pairs of frames (D. Fleet & Weiss, 2006). The displacement vector is known as optical flow vector, expresses in terms of direction and magnitude. For example, Figure 2.9 illustrates the optical flow estimation of the moving object between adjacent frames.

Conventionally, there are several approaches to accomplish the optical flow computation. They are broadly classified into the following four main types: (a) differential (i.e., Horn and Schunk (Horn & Schunck, 1981), Lucas and Kanade (Lucas & Kanade, 1981), Nagel (Nagel, 1983)); (b) region-based matching (i.e., Anandan (Anandan, 1987), Singh (Singh, 1990)); (c) energy-based (i.e., Heeger (Heeger, 1987)), and; (d) phase-based method (i.e., Fleet and Jepson (D. J. Fleet & Jepson, 1990)). To investigate and discover the capability of these four types of optical flow, Barron et al. (Barron et al., 1994) conducted the experiments by applying them in multiple video sequences. It is reported that the local differential method achieved the best performance over the other

three methods. As a result, *differential method* is opted to approximate the optical flow motion vectors in this thesis. The following descriptions and derivations of optical flow are discussed based on the *differential method*.

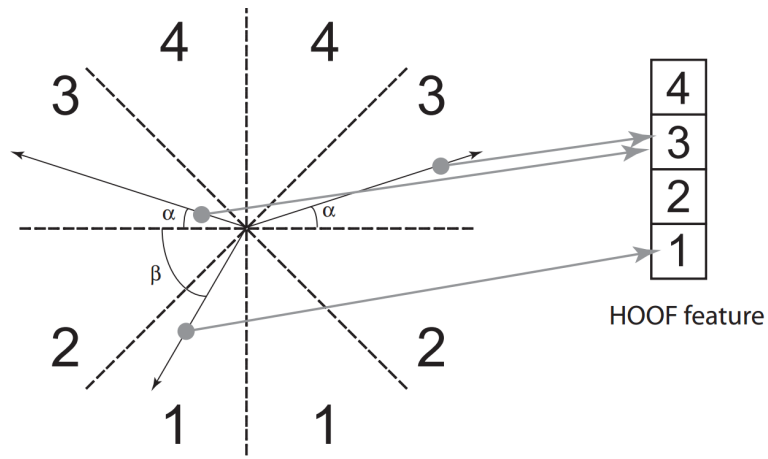
In general, five assumptions are made in order to estimate the optical flow: (a) brightness constancy - apparent brightness of the moving objects remains constant between frames as the changes in the pixel intensity are due solely to motion (highlights, shadows, variable illumination and surface translucency phenomena exists in the images are ignored); (b) object rigidity - the objects in the scene are rigid and the shape changes are neglected; (c) temporal persistence - motion of a surface patch changes gradually over time because the movement between two adjacent frames are small; (d) spatial coherence - the neighborhood of the object are likely to be situated on the same surface. Thus, the neighboring points of the object are having similar velocity values, and; (e) continuous and differentiable - image velocity are approximated from spatio-temporal derivatives of image intensities by using differential method. Therefore, image flow field is continuous and differentiable in both the space and time domains (Beauchemin & Barron, 1995).

The optical flow constraint equation is formulated as:

$$\nabla I \bullet \rho + I_t = 0, \quad (2.8)$$

where  $I(x,y,t)$  represents the changes of temporal image brightness in intensity values with respect to time at point  $(x,y)$ .  $\nabla I = (I_x, I_y)$  is the spatial gradient and  $I_t$  is the temporal gradient of the intensity functions. Assume that the point of interest in the image is initially positioned at  $(x,y)$ . After a change in time by  $dt$ , it moves through a distance  $(dx, dy)$ . The horizontal and vertical components of the optical flow are denoted as:

$$\rho = [p = \frac{dx}{dt}, q = \frac{dy}{dt}]^T. \quad (2.9)$$



**Figure 2.10:** Formation of HOOF with four bins

Optical flow has been employed in a variety of interesting applications, such as facial expression recognition (Kenji, 1991), human skeleton tracking (Schwarz et al., 2012), medical imaging (Garcia et al., 2012), vehicle detection (Dawood et al., 2013), object removal (Hamilton & Breckon, 2016) and many more. In addition, several variants of optical flow have been designed and optimized to best meet different research objectives.

Histogram of Oriented Optical Flow (HOOF) (Chaudhry et al., 2009) is one of the techniques developed based on the optical flow. It was first introduced to recognize human actions by building activity profiles for each frame of a video. The optical flow values are captured as histogram bins because they are extremely vulnerable to the background changes and illumination variation. By doing so, it allows the feature representation to be independent to the direction of the motion and scale variation. To generate HOOF, optical flow for each pair of adjacent frames is first computed. Then, the orientation and magnitude are calculated based on the horizontal and vertical flow vectors. Lastly, the histogram is formed according to the orientation, with magnitude being used as the weight to highlight the importance of each optical flow. Figure 2.10 shows the formation of HOOF with bin number of four.

### 2.3.4 Optical Strain

Optical strain is the extension of optical flow (Gibson, 1950). In contrast to optical flow, optical strain is robust to lighting condition, heavy makeup and camouflage (Shreve et al., 2010; Manohar et al., 2007). Strain is the deformation of an object due to force or stress, while deformation refers to the motion causing the alteration of surface or volume of an object (Heimdal et al., 1998). For a small enough facial pixel movement, it is able to approximate the deformation intensity. As such, it is also sometimes called the infinitesimal strain tensor (Lee, 1969). In brief, infinitesimal strain tensor is derived from Lagrangian and Eulerian strain tensors after performing a geometric linearisation with the assumption of little deformation (Simof & Hughes, 2008). Optical strain can be utilized to measure the intensity of the expression occurred and is therefore useful in facial expression detection or recognition task. Figure 2.11 shows the visualization of optical flow and optical strain computed between onset and apex frames. Here, the occurrence of the micro-expression is called the onset, while the disappearance of the AU is called the offset. On the other hand, apex is the instant when the expression reaches its peak intensity. It is observed that optical strain image is able to highlight more precise facial muscle changes (viz., on the eyes and eyebrow regions).

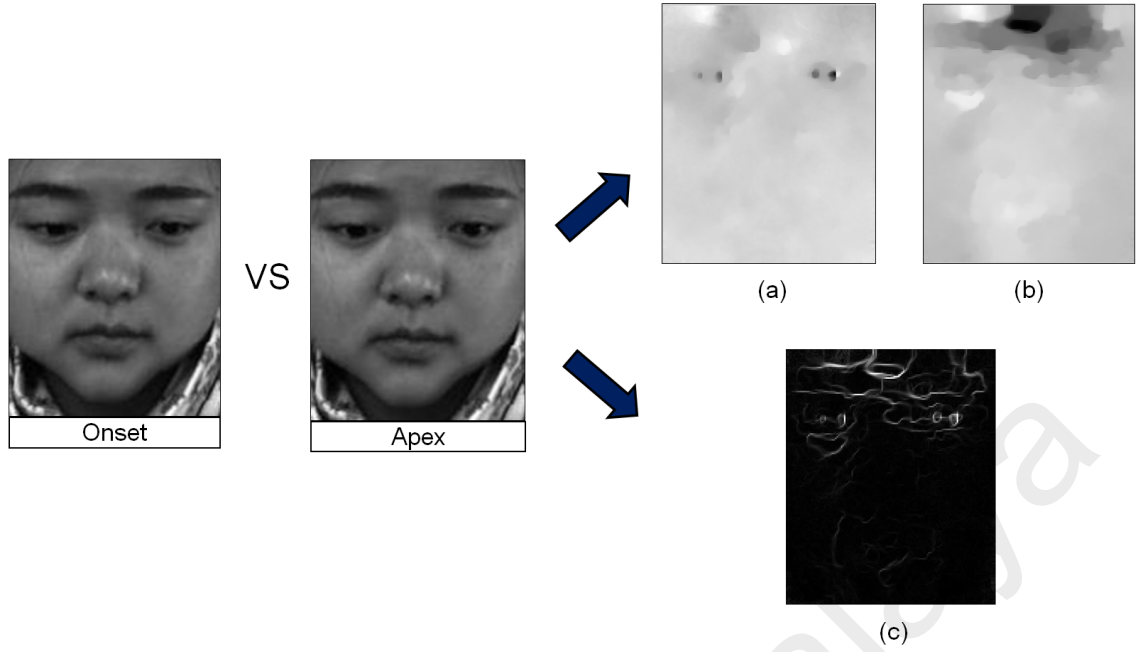
In terms of displacements, a typical infinitesimal strain  $\epsilon$ , can be defined as:

$$\epsilon = \frac{1}{2}[\nabla \vec{u} + (\nabla \vec{u})^T], \quad (2.10)$$

where  $\vec{u} = [u, v]^T$  is the displacement vector. It can also be re-written as:

$$\epsilon = \begin{bmatrix} \epsilon_{xx} = \frac{\partial u}{\partial x} & \epsilon_{xy} = \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) \\ \epsilon_{yx} = \frac{1}{2}(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}) & \epsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix}, \quad (2.11)$$

where the diagonal strain components  $(\epsilon_{xx}, \epsilon_{yy})$  are normal strain components and  $(\epsilon_{xy}, \epsilon_{yx})$



**Figure 2.11:** Optical flow and optical strain computed between the onset and apex frames. Visualization of (a) Horizontal optical flow; (b) Vertical optical flow; (c) Optical strain

are shear strain components. Normal strain measures the changes in length along a specific direction, whereas shear strain measures changes in two angular directions that form the plane experiencing the shear distortion (Yamaji, 2007).

To estimate the strain from the optical strain magnitude (Equation (2.10)), the optical flow vectors  $(p, q)$  in Equation (2.8) can be simplified by differentiating it to the first order derivatives because the strain components are described in a function of displacement vectors  $\vec{u} = [u, v]^T$ . Specifically,

$$p = \frac{dx}{dt} \approx \frac{\Delta x}{\Delta t} = \frac{u}{\Delta t} \implies u = p\Delta t, \quad (2.12)$$

and

$$q = \frac{dy}{dt} \approx \frac{\Delta y}{\Delta t} = \frac{v}{\Delta t} \implies v = q\Delta t, \quad (2.13)$$

where  $\Delta t$  is the time interval between two image frames. Since the temporal resolution of a video is constant,  $\Delta t$  is also a constant value (i.e., fixed length), and the partial

derivative of Equation (2.12) and Equation (2.13) are approximated as:

$$\begin{aligned}\frac{\partial u}{\partial x} &\approx \frac{\partial p}{\partial x} \Delta t, & \frac{\partial u}{\partial y} &\approx \frac{\partial p}{\partial y} \Delta t, \\ \frac{\partial v}{\partial x} &\approx \frac{\partial q}{\partial x} \Delta t, & \frac{\partial v}{\partial y} &\approx \frac{\partial q}{\partial y} \Delta t.\end{aligned}\tag{2.14}$$

The second order derivatives are approximated by using Finite Difference Approximation.

$$\begin{aligned}\frac{\partial u}{\partial x} &= \frac{u(x + \Delta x) - u(x - \Delta x)}{2\Delta x} = \frac{p(x + \Delta x) - p(x - \Delta x)}{2\Delta x}, \\ \frac{\partial v}{\partial y} &= \frac{v(y + \Delta y) - v(y - \Delta y)}{2\Delta y} = \frac{q(y + \Delta y) - q(y - \Delta y)}{2\Delta y}, \\ \frac{\partial u}{\partial y} &= \frac{u(y + \Delta y) - u(y - \Delta y)}{2\Delta y} = \frac{p(y + \Delta y) - p(y - \Delta y)}{2\Delta y}, \\ \frac{\partial v}{\partial x} &= \frac{v(x + \Delta x) - v(x - \Delta x)}{2\Delta x} = \frac{q(x + \Delta x) - q(x - \Delta x)}{2\Delta x},\end{aligned}\tag{2.15}$$

where  $(\Delta x, \Delta y)$  are preset distances of 1 pixel.

The optical strain magnitude for each pixel can be calculated by taking the sum of squares of the normal and shear strain components, expressed as follows:

$$\begin{aligned}|\epsilon_{x,y}| &= \sqrt{\epsilon_{xx}^2 + \epsilon_{yy}^2 + \epsilon_{xy}^2 + \epsilon_{yx}^2} \\ &= \sqrt{\frac{\partial u^2}{\partial x} + \frac{\partial v^2}{\partial y} + \frac{1}{2} \left( \frac{\partial u}{\partial x} + \frac{\partial u}{\partial x} \right)^2}.\end{aligned}\tag{2.16}$$

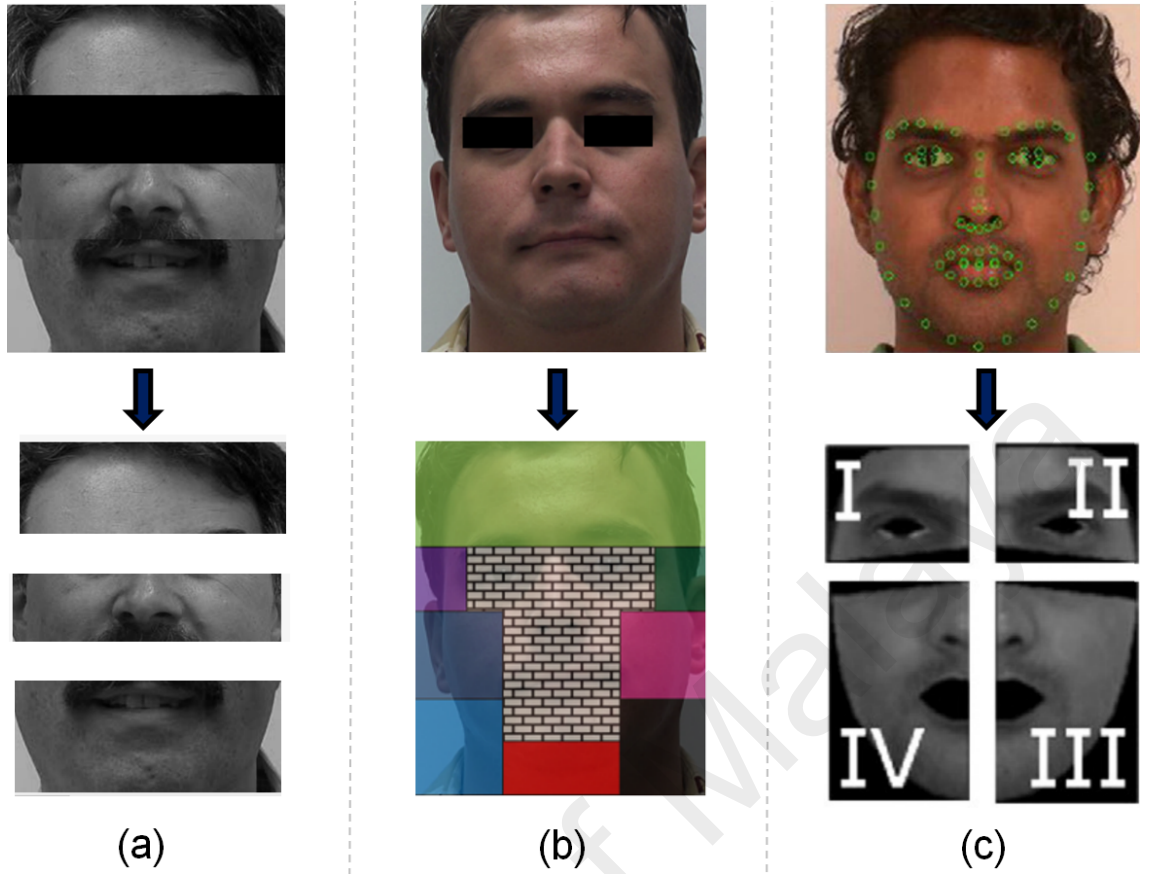
Optical strain has demonstrated its superiority in the literature, especially in the field of micro-expression detection and recognition tasks. At this juncture, it is important to note that *spotting* or *detection* refers to the task of identifying the presence of facial expression (macro- or micro-) without determining the specific type of expression, Whereas



*recognition* or *categorization* is to classify the type of the expression. The work by Shreve (Shreve et al., 2009) outperformed optical flow in micro-expression detection by attaining 100% detection accuracy with a single false spot. The experiments are conducted on the USF (Yan, Wang, Liu, et al., 2014) dataset. It is achieved by spotting the peak of the strain magnitude graph with respect to each face regions (i.e., forehead, cheeks, and mouth) plotted in time series. The example of the face partition is shown in Figure 2.12(a). However, the drawbacks of the experiment is that, the dataset contains only seven samples, which is relatively small. Besides, the micro-expressions spotted are posed rather than spontaneous stimulation. Thus, the spotting task may be easier to be performed as the spotted expressions contain larger and more obvious facial motion.

Few years later, the same authors (Shreve, Godavarthy, et al., 2011) carried out a more extensive test on two more datasets. An improved algorithm is implemented to spot the micro-expressions. For the existing USF-HD database (Shreve, Godavarthy, et al., 2011), the micro-expressions are increased to 100 samples. The two extra datasets are Canal-9 dataset (Vinciarelli et al., 2009) and “found videos” dataset (Ekman, 2009b). The former contains 24 micro-expressions, while the latter consists of 4 micro-expressions. The modified algorithm is able to distinguish between universal macro-expressions and rapid micro-expressions. To capture the local gradient of motion, the face is divided into eight smaller regions, particularly: forehead, left and right of eye, left and right of cheek, left and right of mouth and chin. The illustration of the facial segmentation is shown in Figure 2.12(b). Finally, the peaks in the strain plots (which are obtained from the optical strain magnitudes for each region) are identified as the spotted expressions. A promising spotting accuracy of 74% is obtained.

The latest work by the same group of author Shreve et al. (Shreve et al., 2014) is to examine an enhanced technique to segment out the macro- and micro-expressions frames in video sequences. The mouth and eyes regions of the face are masked because there



**Figure 2.12:** The eyes are masked for privacy concerns. Face is segmented into: (a) Three regions (i.e., forehead, cheeks, and mouth) (Shreve et al., 2009); (b) Eight regions (i.e., forehead, left and right of eye, left and right of cheek, left and right of mouth and chin) (Shreve, Godavarthy, et al., 2011); (c) Four regions (i.e., upper left, lower left, upper right and lower right) (Shreve et al., 2014)

are noises due to violation of the smoothness constraints and self-occlusions. Instead of partitioning the face into eight regions (Shreve, Godavarthy, et al., 2011), they divide the face into four regions: upper left, lower left, upper right and lower right, as shown in Figure 2.12(c). The databases considered to evaluate the proposed method are USF-combination (Shreve et al., 2014) (consists of 37 feigned micro-expression samples) and SMIC micro (Pfister et al., 2011) (contains 77 micro-expression sequences). The best performance obtained is nearly 80% true positive rate for spotting micro-expressions with a 0.3% false positive rate on USF-combination dataset.

Optical strain technique is found to be useful in medical analysis (Shreve, Jain, et al., 2011). Owing to the fact that optical strain highlights subtle changes between two images, it can provide a useful measure of the asymmetries at the precise position on

the face. Therefore, it can potentially allow surgeons to estimate the severity of facial paralysis in a short period instead of marking them manually, which is time consuming and requires a great degree of human effort.

## **2.4 Overview of Micro-expression Databases**

Normal expressions have well-established databases and are being studied comprehensively. In contrast, there are relatively few micro-expression databases in the literature. As a result, the new micro-expression techniques developed have limited databases to be evaluated and analyzed on, thus hindering the progress in related researches. According to the benchmark of micro-expression defined by Ekman (Ekman & Friesen, 1969), micro-expression has to be both micro and subtle. In this section, the existing micro-expression databases and the problems faced by them are discussed. Each database is elaborated in each sub-chapter, namely SMIC (Pfister et al., 2011), SMIC II (Li et al., 2013), CASME (Yan et al., 2013), CASME II (Yan, Li, et al., 2014), and others. Note that the videos collected are unequally distributed in different classes due to the difficulties in eliciting some particular types of the micro-expressions.

### **2.4.1 SMIC**

Spontaneous Micro-expression (SMIC) dataset is made up of 77 spontaneous and dynamic videos from six participants (three males and three females). The videos are recorded using PixelINK PL-B774U camera with a temporal resolution of 100 frames per second (fps) and spatial resolution of  $640 \times 480$  pixels. The average frame length is 29 frames ( $\sim 0.3$  s) and the shortest expression is 11 frames ( $\sim 0.11$  s). The videos are elicited by asking the participants to maintain a neutral face and suppress their genuine feeling whilst watching the clips. The researchers attempt to guess the type of the video clip episode the participants are watching. If the researchers guess it correctly, the participants are asked to fill in a long and dull questionnaire. The collected video record-

ings are then labeled by two coders, whereby the clips are first viewed frame by frame before repeating the viewing process with increasing speed. This ground-truth labeling process follows the advice proposed by Ekman (Ekman, 2009a). Lastly, the expression categories are determined by two coders after considering the self-reported emotions by the participants.

To establish the baseline performance, all the recorded micro-expression videos undergo three main processes (i.e., pre-processing, feature extraction and classification) to obtain the expression recognition results. Firstly, in the pre-processing step, a Haar eye detector (Niu et al., 2006) is adopted to obtain the location of the eyes. Then, ASM (Cootes et al., 1995) is employed to annotate the facial feature points, which are used for face segmentation and normalization. LWM (Goshtasby, 1988) is exploited to transform the feature point from the original face into a pre-defined template face. All the normalized face images are interpolated into a certain frame number by Temporal Interpolation Model (TIM) (Zhou et al., 2012) to address the problem of short video lengths. A simple graphical illustration of TIM is shown in Figure 2.13(b), where a video is mapped onto the curve to generate new video data. Secondly, block-based LBP-TOP (Zhao & Pietikainen, 2007) is employed as feature descriptor with the block size of  $5 \times 5$  and  $8 \times 8$ . Thirdly, three different types of classifier, including: (a) Support Vector Machine (SVM) (Suykens & Vandewalle, 1999) with Leave-One-Subject-Out Cross Validation (LOSOCV) setting; (b) Random Forest (RF) (Breiman, 2001), and; (c) Multiple Kernel Learning (MKL) (Varma & Ray, 2007), are adopted for the detection and recognition tasks. Detection classifies micro-expressions versus non-micro-expressions, whereas recognition classifies the type of emotion (i.e., either negative, positive or surprise). For LOSOCV in SVM, the video sequence of one subject is treated as the testing data and the remaining frames as the training data. This process is repeated for  $k$  times, where  $k$  is the number of subjects in the database. Then, the recognition results for all subjects

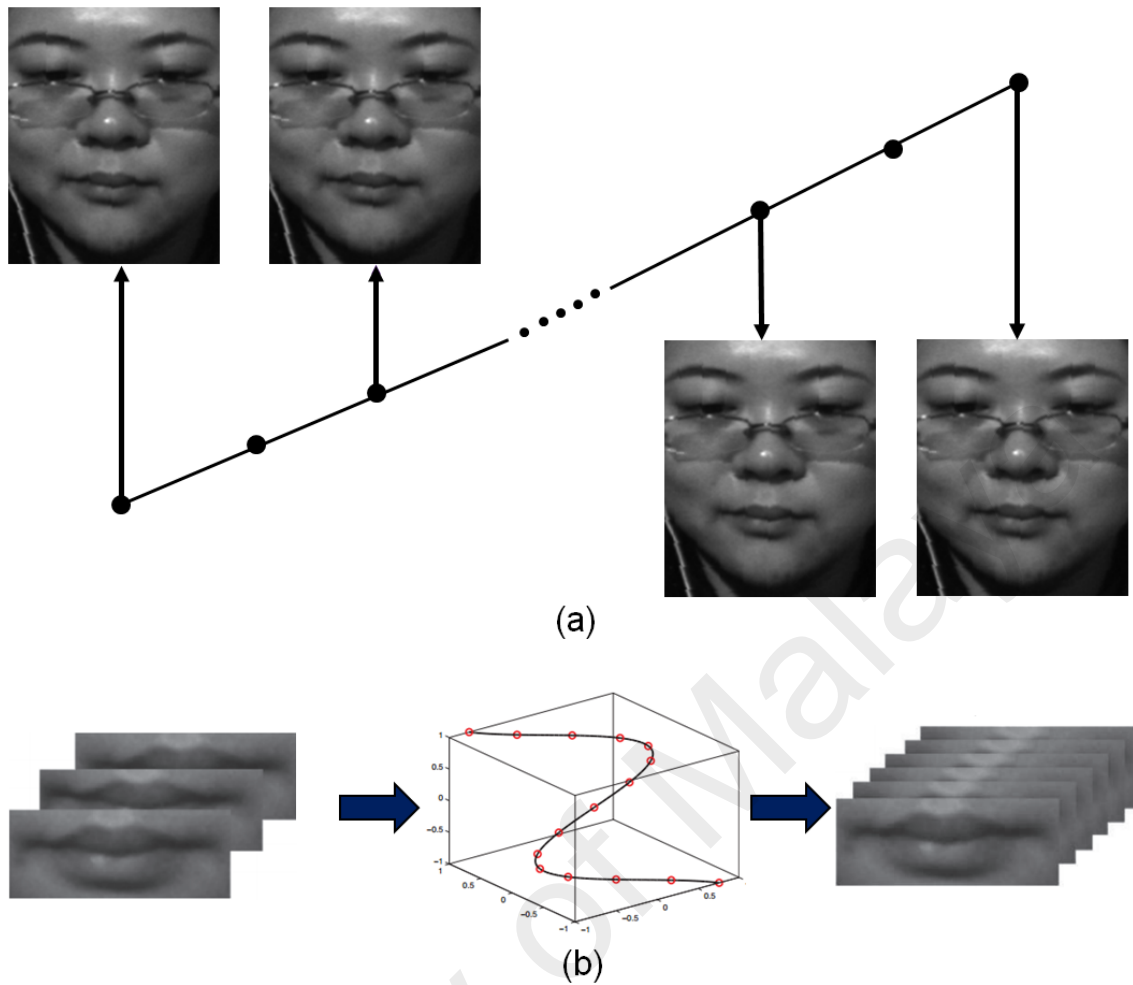
**Table 2.2:** Detailed information of the SMIC database

Participants		SMIC 16
Camera	Type Frame rate (fps)	PixeLINK PL-B774U 100
Image resolution (pixels)		640 × 480
Total Expression		77
Frame number	Average	29
	Minimum	11
Video duration (s)	Average	0.30
	Minimum	0.11
Pre-processing technique		Haar eye detector ASM LWM TIM
Feature Extractor		LBP-TOP
Classifier		SVM RF MKL
Best Result (%)	Detection	74.30
	Recognition	71.40

are averaged to form the final recognition accuracy. The best accuracy results obtained in SVM classifier is 74.3% for the detection task and 71.4% for the recognition task. The data composition of SMIC database is tabulated in detail in Table 2.2.

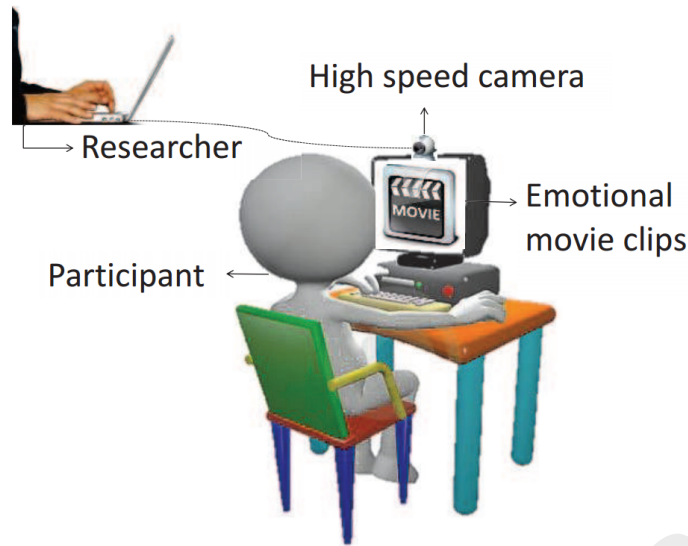
#### 2.4.2 SMIC II

Two years later, a newer Spontaneous Micro-expression (SMIC) dataset is established by the same group of researchers (Li et al., 2013). This database is named as SMIC II to clear up the confusion between this database (Li et al., 2013) and the one with the same name published in 2011 (Pfister et al., 2011). This database contains 164 micro-expression video clips elicited from 16 participants with an average age of 28.1 years. It is made up of six females, ten males, where eight Asians and eight Caucasians are involved. To elicit the micro-expression, the participants are asked to watch several short emotional video clips and to try to maintain their head position while watching the clips. The participants are asked to put on a poker face and to not reveal their true feelings.



**Figure 2.13:** (a) SMIC sample images; (b) Video mapping on the curve by adopting TIM to produce a new video

While the participants are watching the films, the researchers stay in the other room to observe their facial and body movements through the camera, and at the meantime try to guess which clip they are watching. They are given a short break after watching each video clip. If they failed to hide their feelings or, in other words, the researchers guess the correct video clip, the participants are asked to fill in a very long and boring questionnaire of more than 500 questions. The acquisition setup of micro-expressions elicitation for SMIC II is shown in Figure 2.14. The camera is placed right on top of the screen to capture the facial expression of the participants. After collecting the micro-expression videos, the expression categories are determined by two coders based on the participants' self-report data. The ground-truths of the dataset are provided, which include the onset, offset of the expression, the represented expression, and the marked Action Unit (AU).

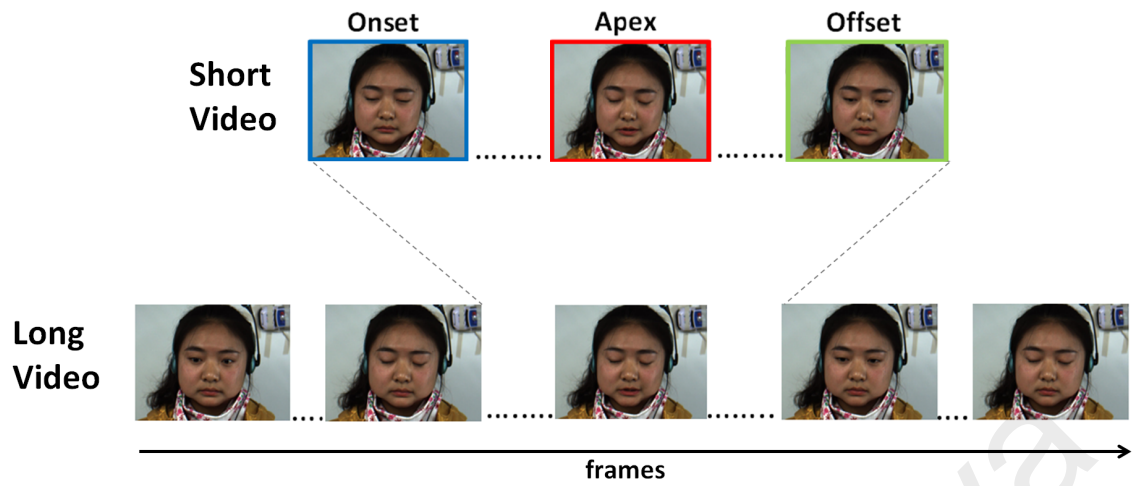


**Figure 2.14:** The acquisition setup for micro-expression elicitation of SMIC II database

SMIC II database consists of two sub-datasets, which are the *long video* and the *short video*. The video sequence that contains only frames from onset to offset is defined as *short video*. On the other hand, *Long video* refers to the raw video sequence which may include the frames with micro-expressions and other irrelevant motions that present before the onset and after the offset. The sub-datasets grouped under the *short video* category are SMIC-HS, SMIC-VIS and SMIC-NIR, which are recorded with different cameras at slightly different positions. The sub-datasets belonging to the *long video* category are SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR. The details of these six sub-datasets are described in the following sub-chapters. Figure 2.15 shows the graphical representation of short and long videos with annotated ground-truth labels, including the onset, apex and offset frames.

#### 2.4.2.1 Short Video

In this sub-chapter, detailed information of the three *short video* datasets in SMIC II are discussed, namely, SMIC-HS, SMIC-VIS and SMIC-NIR. In addition, the baseline detection and recognition performances for each dataset are provided. The pre-processing method exploited in these datasets are ASM and Haar eye detector, while the feature



**Figure 2.15:** Example of short and long videos with onset, apex and offset annotations

descriptor employed is block-based LBP-TOP. SVM with LOSOCV configuration is adopted as the classifier.

Table 2.3 summarizes the general information about short videos in SMIC II database.



**Table 2.3:** Detailed information of the SMIC-HS, SMIC-VIS and SMIC-HR datasets

Participants		SMIC-HS	SMIC-VIS	SMIC-NIR
Camera		16	8	8
Type		PixeLINK PL-B774U	Visual camera	Near-infrared camera
Frame rate (fps)		100	25	25
Image resolution (pixels)		$640 \times 480$	$640 \times 480$	$640 \times 480$
Expression	Positive	51	28	28
	Negative	70	23	23
	Surprise	43	20	20
	Total	164	71	71
Frame number	Average	34	10	10
	Maximum	58	13	13
	Minimum	11	4	4
Video duration (s)	Average	0.34	0.40	0.40
	Maximum	0.58	0.52	0.52
	Minimum	0.11	0.16	0.16
Ground-truth		Onset and offset frame indices		
		Emotion label		
		Action unit label		
Pre-processing technique		Haar eye detector ASM		
Feature Extractor		LBP-TOP		
Classifier		SVM		
Best Result (%)	Detection	65.55	62.68	59.15
	Recognition	48.78	52.11	38.03

## **SMIC-HS**

This dataset is made up of 164 video clips from 16 participants. The videos are recorded using PixeLINK PL-B774U camera with a temporal resolution of 100 fps and spatial resolution of  $640 \times 480$  pixels. The average frame length is 34 frames ( $\sim 0.34$  s); the longest being 58 frames ( $\sim 0.58$  s) and the shortest being 11 frames ( $\sim 0.11$  s). There are three emotion classes: surprise (43 videos), positive (51 videos) and negative (70 videos).

A three-class baseline recognition accuracy is reported as 48.78% by employing block-based LBP-TOP with block size of  $8 \times 8$  as the feature descriptor and TIM of 10 frames. For detection, the accuracy achieved is 65.55% by changing the block size of LBP-TOP to  $5 \times 5$ .

## **SMIC-VIS**

It is a collection of 71 videos obtained from eight subjects. The videos are recorded using a standard visual camera at a resolution of  $640 \times 480$  pixels at 25 fps. The average frame length is 10 frames ( $\sim 0.4$  s); the longest is 13 frames ( $\sim 0.52$  s) and the shortest is 4 frames ( $\sim 0.16$  s). The videos are categorized into three classes, namely: surprise (20 videos), positive (28 videos) and negative (23 videos).

The baseline performance for this three-class recognition task is 52.11%. An LBP-TOP with  $5 \times 5$  block partitioning and TIM of 10 frames are adopted. For the detection task, the block size of LBP-TOP when set to  $5 \times 5$  generates the best result of 62.68%.

## **SMIC-NIR**

It consists of 71 clips from eight subjects. The videos are recorded using a near-infrared camera with a frame rate of 25 fps at  $640 \times 480$  pixels. The video clips have an average length of 10 frames ( $\sim 0.4$  s); the longest being 13 frames ( $\sim 0.52$  s) while the shortest

is 4 frames ( $\sim 0.16$  s). It consists of three classes, namely: surprise (20 videos), positive (28 videos) and negative (23 videos).

A recognition results of 38.03% is obtained by employing LBP-TOP with block partitions of  $8 \times 8$  with TIM of 10 frames. The detection accuracy of 59.15% is achieved when block size is  $8 \times 8$  and TIM is 20 frames.

#### 2.4.2.2 Long Video

In this sub-chapter, detailed information of the three *long video* datasets in SMIC II are discussed, namely, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR. The detailed information for long videos in SMIC II database is tabulated in Table 2.4.

**Table 2.4:** Detailed information of the SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR datasets

Participants		SMIC-E-HS	SMIC-E-VIS	SMIC-E-NIR
Camera		16	8	8
Type		PixeLINK PL-B774U	Visual camera	Near-infrared camera
Frame rate (fps)		100	25	25
Image resolution (pixels)		640 × 480	640 × 480	640 × 480
Expression	Positive	51	28	28
	Negative	66	24	23
	Surprise	40	19	20
	Total	157	71	71
Frame number	Average	590	150	150
	Maximum	1200	300	300
	Minimum	120	30	30
Video duration (s)	Average	6	6	6
	Maximum	12	12	12
	Minimum	1.2	1.2	1.2
Ground-truth		Onset and offset frame indices		
		Emotion label		
		Action unit label		

### **SMIC-E-HS**

It consists of 157 micro-expression clips from 16 subjects (mean age of 28.1 years). The videos are recorded using PixeLINK PL-B774U camera with a temporal resolution of 100 fps and spatial resolution of  $640 \times 480$  pixels. The average frame length is 590 frames ( $\sim 6$  s); the longest being 1200 frames ( $\sim 12$  s) while the shortest is 120 frames ( $\sim 1.2$  s). There are three micro-expression classes, including: negative (66 videos), positive (51 videos) and surprise (40 videos).

### **SMIC-E-VIS**

This dataset is made up of 71 micro-expression videos from eight subjects. The videos are recorded using a standard visual camera with a frame rate of 25 fps at  $640 \times 480$  pixels. The video clips have an average length of 150 frames ( $\sim 6$  s); the longest being 300 frames ( $\sim 12$  s) and the shortest is 30 frames ( $\sim 1.2$  s). It consists of three micro-expression classes, including: negative (24 videos), positive (28 videos) and surprise (19 videos).

### **SMIC-E-NIR**

It is a collection of 71 micro-expression video sequence obtained from eight subjects. The videos are recorded with a near-infrared camera at a resolution of  $640 \times 480$  pixels at 25 fps. The average frame length is 150 frames ( $\sim 6$  s); the longest being 300 frames ( $\sim 12$  s) and the shortest is 30 frames ( $\sim 1.2$  s). The micro-expression videos are categorized into three classes, including: negative (23 videos), positive (28 videos) and surprise (20 videos).

### 2.4.3 CASME

The Chinese Academy of Sciences Micro-expression (CASME) database comprises of 195 micro-expressions from 35 subjects (13 females, 22 males) with an average age of 22.03 years and standard deviation of 1.6. There are eight main emotion classes of micro-expression, in the following distribution: 2 fear videos, 3 contempt videos, 5 amusement videos, 6 sadness videos, 20 surprise videos, 28 tense videos, 40 repression videos, and 88 disgust videos. The micro-expressions are elicited from the participants by showing them the video episodes downloaded from the Internet. During the recording period, they are asked to maintain a neutral face, while keeping their eyes still and their head stationary. To enhance the elicitation process, the participants are not allowed to show any facial expressions, or else their remuneration will be deducted. The emotion types are labeled based on the participants' self-report and two psychological researchers. A reliability score of the AU labeling is reported at 0.83. The ground-truths provided include the onset, apex, offset frames indices, the AUs labels as well as the emotion classes.

There are two sub-categories in this database (i.e., CASME A and CASME B) that are recorded under different environmental configuration and with different camera type. More precisely, in CASME A, the videos are collected using BenQ M31 camera at a spatial resolution of  $1280 \times 720$  pixels and a frame rate of 60 fps. The average total duration is 289.96 ms ( $\sim 17$  frames), with the average onset duration is 130.32 ms ( $\sim 7$  frames). For CASME B, the videos are recorded with a Point Grey GRAS-03K2C camera at a resolution of  $640 \times 480$  pixels and frame rate of 60 fps. The video clips have an average total duration of 299.24 ms ( $\sim 18$  frames) and an average onset duration of 123.99 ms ( $\sim 7$  frames).

The contents of CASME A and CASME B databases are summarized and tabulated in Table 2.5.

**Table 2.5:** Detailed information of CASME A and CASME B databases

		CASME A	CASME B
Participants		35	
Camera	Type Frame rate (fps)	BenQ M31 60	Point Grey GRAS-03K2C 60
Image resolution (pixels)		1280 × 720	640 × 480
Expression	Fear	2	
	Contempt	3	
	Amusement	5	
	Sadness	6	
	Surprise	20	
	Tense	28	
	Repression	40	
	Disgust	88	
Frame number	Average	17	18
	Onset	7	7
Video duration (s)	Average	0.3	0.3
	Onset	0.1	0.1
Ground-truth		Onset and offset frame indices Emotion label Action unit label	
Reliability score		0.83	

#### 2.4.4 CASME II

Subsequent to CASME (Yan et al., 2013) database development, the same group of researchers (Yan, Li, et al., 2014) published a newer database with more micro-expression samples. Similar to SMIC II (refer to Chapter 2.4.2), it consists of two sub-datasets (i.e., *short videos* and *long videos*). To clarify the difference between the *short videos* and *long videos* in this dissertation, the dataset for *short videos* is named as CASME II, while the dataset for the *long videos* is named as CASME II-RAW. These databases are recorded using a 200 fps high speed camera, Point Grey GRAS-03K2C, with the image resolution set to 640 × 480. 26 participants with an average age of 22.03 years and standard deviation of 1.6 are involved in the databases. The elicitation of the micro-expression videos is similar to that in CASME (Yan et al., 2013), except that several new video episodes that are used to trigger participants' expressions are added and some of the old ones are removed. Besides, the acquisition environment and setup are improved for the record-

ing process. As illustrated in Figure 2.16, four LED lamps and umbrella reflectors are meticulously positioned to reduce the illumination variation on the face and to prevent the flickering light from being captured by the camera. The emotion labels are annotated based on the action unit labels marked by two coders, participants' self report and the content of the video episodes. The reliability score for the action unit labeling performed by the coders is 0.85. The ground-truths provided include onset, apex, offset frame indices, action unit labels and the emotion classes. Information for CASME II and CASME II-RAW is elaborated in the following sub-chapters and is summarized in Table 2.6.

#### 2.4.4.1 CASME II

This dataset consists of 247 micro-expression video clips, particularly, 25 surprise videos, 27 repression videos, 33 happiness videos, 60 disgust videos, and 102 others videos. The video clips have an average length of 68 frames ( $\sim 0.3$  s); the longest being 141 frames ( $\sim 0.7$  s) and the shortest is 24 frames ( $\sim 0.1$  s). To establish a baseline recognition performance,  $5 \times 5$  block partitioning in LBP-TOP is adopted. The features are classified by SVM with Leave-One-Video-Out Cross Validation (LOVOCV) protocol. For LOVOCV principle, only one video sequence is treated as testing data and the remainder as training data. This process is repeated for  $k$  times and the results are averaged to obtain the resultant recognition accuracy, where  $k$  is the total number of videos in the database. A five-class classification of 63.41% is reported.

#### 2.4.4.2 CASME II-RAW

This dataset is a collection of 246 micro-expressions videos. The distribution of the expression videos is: surprise (25 videos), repression (27 videos), happiness (32 videos), disgust (63 videos), and others (99 videos). The average frame length is 244 frames ( $\sim 1.22$  s), with the longest being 1,024 frames ( $\sim 5.12$  s) and the shortest is 51 frames ( $\sim 0.26$  s).





**Figure 2.16:** The acquisition setup for micro-expression elicitation of CASME II database

#### 2.4.5 Other Micro-expression Databases

Apart from the aforementioned datasets, there are also other micro-expression databases adopted in developing detection and recognition algorithms. This sub-chapter briefly describes these databases and points out their problems. Specifically, these databases include: (a) USF-HD (Shreve, Godavarthy, et al., 2011); (b) Polikovsky's (Polikovsky et al., 2009), and; (c) YorkDDT (Warren et al., 2009; Pfister et al., 2011). The general information of these three databases are tabulated in Table 2.7.

USF-HD database contains 100 micro-expressions with six different emotion classes. Participants are asked to perform macro and micro-expressions and the videos are recorded at the frame rate of 30 fps with a resolution of  $720 \times 1280$ . The main drawback of this database is that they are posed micro-expressions rather than spontaneous ones, which is against the principle of micro-expression. Besides, the micro-expression videos have longer duration (two-third of a second), far exceeding the normal definition of one-fifth of a second.

In Polikovsky's database, 10 participants are asked to perform seven basic emotions with low facial muscle intensity so that they can go back to the neutral face expression as

**Table 2.6:** Detailed information of the CASME II and CASME II-RAW databases

		CASME II	CASME II-RAW
Participants		26	
Camera	Type Frame rate (fps)	Point Grey GRAS-03K2C 200	
Image resolution (pixels)		640 × 480	
Expression	Surprise	25	25
	Repression	27	27
	Happiness	33	32
	Disgust	60	63
	Others	102	99
	Total	247	246
Frame number	Average	68	244
	Maximum	141	1024
	Minimum	24	51
Video duration (s)	Average	0.3	1.2
	Maximum	0.7	5.1
	Minimum	0.1	0.3
Ground-truth		Onset, apex and offset frame indices Emotion label Action unit label	
Reliability score		0.85	
Pre-processing technique		ASM LWM	N/A
Feature Extractor		LBP-TOP	N/A
Classifier		SVM	N/A
Best result (%)	Recognition	63.41	N/A

**Table 2.7:** General information of the USF-HD, Polikovsky's and YorkDDT databases

		USF-HD	Polikovsky's	YorkDDT
Participants		N/A	10	9
Camera	Frame rate (fps)	300	200	N/A
Image resolution (pixels)		720 × 1280	480 × 640	320 × 240
Expression	Type	6	6	N/A
	Total	100	42	18
Posed/ Spontaneous		Posed	Posed	Spontaneous

fast as possible. The videos are recorded under a frame rate of 200 fps with a resolution of 480 × 640. A total of 42 micro-expression samples are collected with six different emotion categories. Again, the drawback is that they are posed micro-expressions rather than spontaneous one.

In YorkDDT database, there are 18 micro-expressions elicited from nine participants

(3 males and 3 females). Among the micro-expressions, 7 are from emotional (i.e., truthful) and 11 are from non-emotional (deception) scenarios. The videos are recorded with a resolution of  $320 \times 240$  pixels. The micro-expressions are spontaneous but incorporated with other irrelevant facial movements. Besides that, the sample size is small, which is insufficient for proper experiment.

## 2.5 Summary and Limitations

An overview of a typical micro-expression recognition system was presented, including pre-processing and feature extraction stages. For the pre-processing stage, face registration and alignment steps are the most common processes applied on the datasets. This is because irrelevant motions that are larger than the micro-expression movements (i.e., head movements) can greatly affect the recognition performance. Due to the fact that the state-of-the-art landmark annotators are incapable of detecting the exact locations of the facial feature points for each frame, only the first frame of the videos are registered based on the model face image, while the remaining frames would re-use the transformation matrix derived from the first frame. The second pre-processing technique is image filtering. Filtering process blurs the face image and minimizes the noises caused by changes in illumination or light flickering effects. A variety of filters are established in the literature but the images in different databases may involve different kinds of noise. Thus, a data driven approach is necessary to determine the ideal filter type. The third pre-processing procedure is to select facial regions which generate the most significant and meaningful features that best represent the expressions. Nonetheless, to date, there is no perfect or standard facial patches combination (i.e., size, shape and position) in obtaining good facial expression recognition results. The RoI extraction technique suggested in the conventional work are evaluated on different databases which incorporate various types of nature and attribute.

The next stage in a micro-expression recognition system is the feature extraction process. The two feature extractors, namely, LBP and LBP-TOP, were discussed in the earlier sub-chapters. These two feature extractors are highly sensitive to the noises. Any noises in the image can cause a large impact to the output features. Due to the thresholding scheme of the feature descriptors, they may not operate consistently on the areas with constant gray level. On the other hand, optical flow and optical strain feature descriptors are computationally intensive due to the complexity of the derivative operations. Thus, a real-time application by implementing optical flow or optical strain is not possible. Besides, noises in the image poses a great disadvantage againsts optical flow estimation, as optical flow operates by searching the matching pixels between a pair of frames.

In order to evaluate the methods proposed in this study, it is essential to conduct the experiments on several micro-expression databases. However, to date, there are only a few micro-expression databases that are eligible for training and testing purposes. The first requirement on the database is that it has to have a large number of samples to verify the robustness of the proposed method. Secondly, a camera with high frame rate is required in order to record sufficient number of frames. Thirdly, a proper acquisition setup for expression elicitation is crucial to ensure that clear facial movements are recorded, and to avoid capturing unnecessary noises at the same time. However, there are some limitations in the current databases. One of the shortcomings is the inconsistency among the participants when interpreting the video clips displayed on the screen. For instance, a participant may think of a scene of chewing worms as funny while others may feel disgusting. Furthermore, the video elicitation process is always performed in one specific constrained laboratory condition, yet the elicited expression may be altered in different environment. Lastly, most participants have ages between 20 to 30 years. A wider range of ages may benefit the analysis of the study. Among all the micro-expression databases discussed above, CASME II and SMIC-HS are the most valid databases to evaluate the

proposed methods. This is because the micro-expressions in these two databases are spontaneous, rapid, subtle and contain sufficient number of sample size.

University of Malaya

## CHAPTER 3: HYBRID FACIAL REGIONS SELECTION

### 3.1 Overview

This chapter presents in detail a proposed pre-processing method. Due to the subtlety and brief duration of the micro-expression occurrence on the face, it is essential to select particular expressive and meaningful facial areas for efficient recognition of micro-expressions. Specifically, a facial region selection technique is introduced, namely *RoI-Selective*, which is a hybrid of both heuristic-based and automatic approaches. The heuristic-based identification of important facial regions provides the statistics of the occurrence frequency for the facial action units in the micro-expression videos, while the automatic detection of the landmark points are performed using the state-of-the-art Discriminative Response Map Fitting (DRMF) method. As a consequent of the fusion of the two approaches, three essential Regions of Interest (RoIs) are formed. Two spatio-temporal feature extractors, namely, OSF and block-based LBP-TOP, are applied to describe the local textural and motion features in each facial RoI. Experiments on two recent spontaneous micro-expression databases confirm the effectiveness of considering only the most salient facial regions for the purpose of recognizing micro-expressions. A thorough analysis highlights the range of region-tuning parameters that generates optimum results, as well as computational savings offered by the proposed method. Results suggest that the proposed method is appropriate for micro-expression analysis.

### 3.2 Motivation

Until now, researches on micro-expression recognition are sparsely found. Among the very few methods proposed: (a) Polikovsky et al. (Polikovsky et al., 2009) employed a 3D-gradient descriptor; (b) Wang et al. (S. J. Wang, Chen, et al., 2014) characterized a gray-scale video clip of micro-expression as a 3rd-order tensor, using Discriminant

Tensor Subspace Analysis (DTSA) (S. J. Wang et al., 2011) for feature description and Extreme Learning Machine (ELM) (G. B. Huang et al., 2004) as classifier; (c) Wang et al. (S. J. Wang, Yan, et al., 2014) extract Local Spatio-temporal Directional Features (LSTD) features (Zhao & Pietikäinen, 2013), and; (d) Pfister et al. (Pfister et al., 2011) utilized a TIM to normalize the frame numbers of micro-expression videos, before applying LBP-TOP descriptor (Zhao & Pietikainen, 2007) to obtain features. All methods utilize the entire facial area for feature extraction, with the exception of the work by (c) Wang et al. (S. J. Wang, Yan, et al., 2014). This is seemingly logical but it is hypothesized that those facial regions which typically do not contribute to micro-expressions might instead, be more susceptible towards noise or minor variations at the pixel level. Therefore, are we looking at where it matters?

### **3.3 Literature Review**

#### **3.3.1 Action Unit**






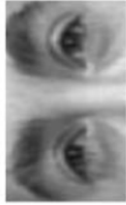






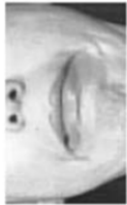





In 1978 (Ekman & Friesen, 1978), Ekman and Friesen established Facial Action Coding System (FACS) to determine the relation between the facial muscle changes and the emotion state. FACS can be used to identify the exact time spot of the beginning and ending for each Action Unit (AU), where AUs are elementary components of FACS representing the action of individual muscles or a cluster of muscles. In 2002, Ekman et al. (Ekman et al., 2002) revised and updated the FACS to 46 AUs, of which 18 cover the lower face and 12 are for the upper face. Figure 3.1 shows some AUs of the FACS with their interpretations. The sample images are excerpted from a macro-expression database – CK+ (Lucey et al., 2010), thus the facial motions are obvious.

The full temporal pattern of the facial expression occurrence is in a sequential order of neutral-onset-apex-offset-neutral (M. F. Valstar & Pantic, 2007). Specifically, onset is the span between the first visible AU to the apex of the AU, while offset is defined as the

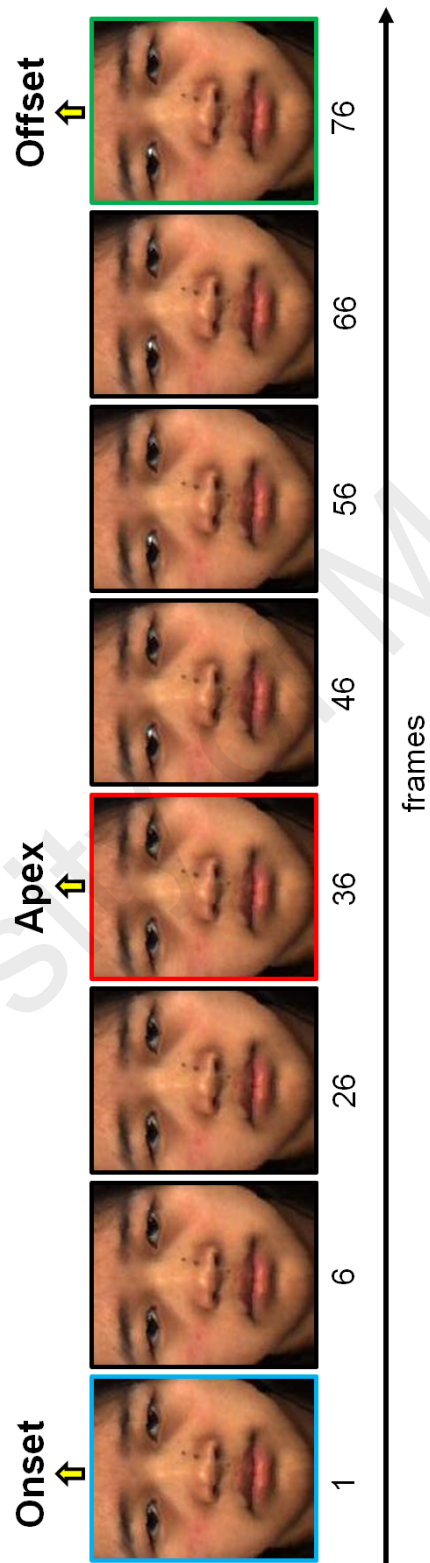
span from the end of the apex of the AU until the disappearance of the AU. On the other hand, apex frame is the location where the AU reaches the peak or the highest intensity of the facial motion. Noted that the location of the onset, offset and apex for the AUs may be different for the same emotion. Figure 3.2 illustrates a sample video annotating the onset, apex and offset frame indices, that shows a ‘Surprise’ expression, taken from the CASME II database (Yan, Li, et al., 2014).

University of Malaya



<b>AU 1</b>  Inner Eyebrow Raiser	<b>AU 2</b>  Outer Brow Raiser	<b>AU 4</b>  Brow Lowerer	<b>AU 5</b>  Upper Lid Raiser	<b>AU 6</b>  Cheek Raiser	<b>AU 7</b>  Lid Tightener
<b>AU 9</b>  Nose Wrinkler	<b>AU 10</b>  Upper Lip Raiser	<b>AU 12</b>  Lip Corner Puller	<b>AU 15</b>  Lip Corner Depressor	<b>AU 16</b>  Lower Lip Depressor	<b>AU 17</b>  Chin Raiser
<b>AU 20</b>  Lip Stretcher	<b>AU 23</b>  Lip Tightener	<b>AU 24</b>  Lip Pressor	<b>AU 25</b>  Lips Part	<b>AU 26</b>  Jaw Drop	<b>AU 27</b>  Mouth Stretcher

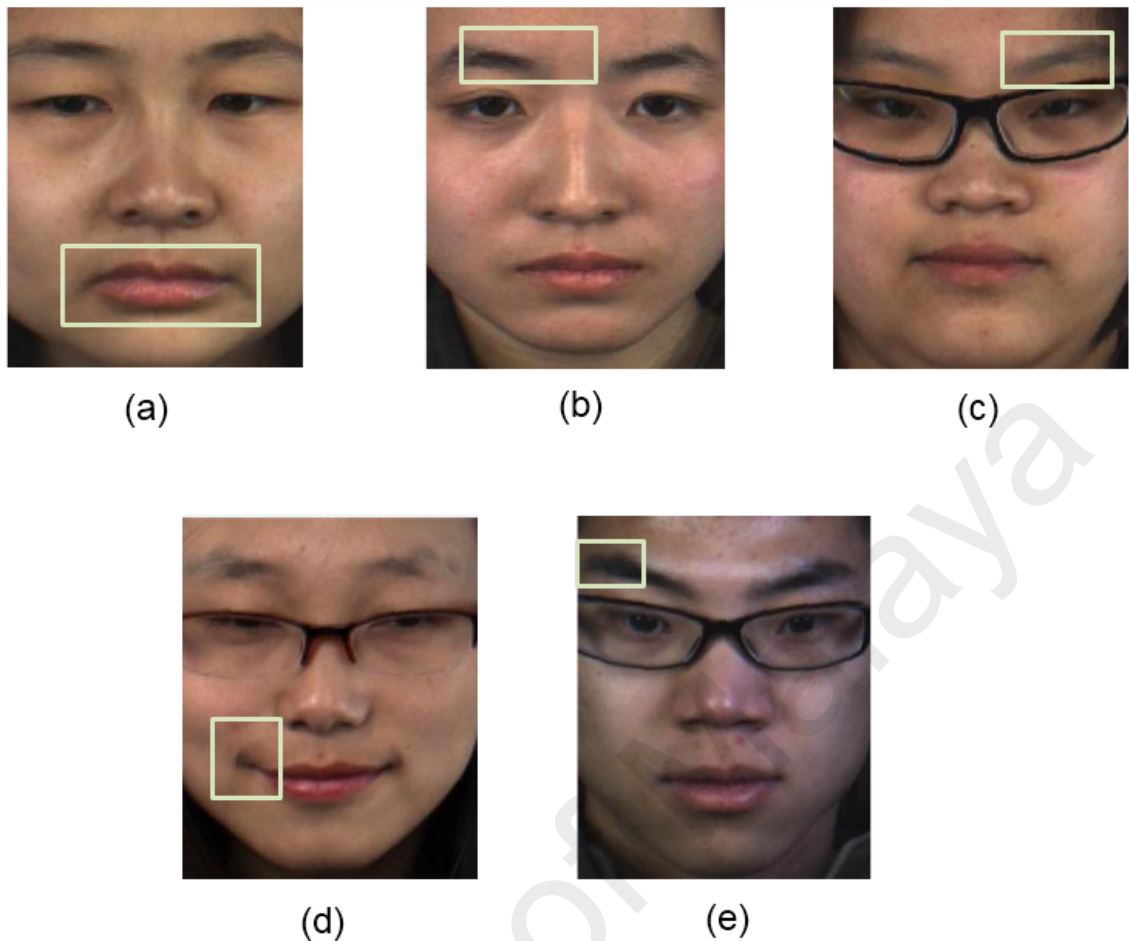
**Figure 3.1:** Example of AUs of the FACS and their interpretations, adapted from (Y. Zhang & Ji, 2005)



**Figure 3.2:** An example of a micro-expression video with onset-apex-offset frame annotation, showing a 'Surprise' expression

FACS expresses the movements of human facial muscles from the appearance. Facial emotion ratings are based on the categorization of expressions into six classical emotions, namely, happiness, surprise, anger, sad, fear, and disgust (Ekman & Friesen, 1971). FACS can describe subtle and ambiguous expressions, thus it is more appropriate to describe the tiny instant changes on the face. Furthermore, an AU represents the facial muscle movements on a specific facial region in single direction. For example, referring to Figure 3.1, AU 23 is lip tightener and AU 24 is lip pressor. AUs can occur either singly or in combination. For instance, both AU 2 and AU 1+2 represent the surprise emotion state, where AU 1 is the inner brow raiser while AU 2 is the outer brow raiser (Yan, Li, et al., 2014). The intensity of the AU can be estimated using a six-point scale, from Neutral < (A) trace < (B) slight < (C) marked/ pronounced < (D) severe/ extreme < (E) maximum.

Facial action unit has been studied intensively, specifically in the analysis on AU detection, recognition and intensity level estimation (M. F. Valstar et al., 2015; Savran et al., 2012; M. Valstar & Pantic, 2006; Rudovic et al., 2014; Jiang et al., 2014), and most recently for micro-expression analysis (S. J. Wang, Yan, et al., 2014) as well. It has been reported that the accuracy of facial expression recognition can be improved with the prior knowledge of the AU information. This is intuitive as AUs are essential for establishing geometrical structure of various facial landmarks. However, a notable challenge with micro-expressions is that the AUs (or a combination of AUs) that are associated with each emotion classes not only differ from that of normal expressions, but are themselves highly variable within each class. For instance, the ‘disgust’ emotion in standard FACS (Ekman et al., 2002) is encoded by AUs 9, 15 and 16, but psychologists in the micro-expression domain have labeled it with AUs 9 or 10 or 4+7 (Yan, Li, et al., 2014), or any combination of these. Figure 3.1 demonstrates the emotions and their AU labels in CASME II database.



**Figure 3.3:** Examples of the emotion and AU labels (facial movements are highlighted) in CASME II database (Yan, Li, et al., 2014) : (a) Fear - AU 20; (b) Sadness - AU 1; (c) Disgust - AU L4; (d) Happiness - AU R12; (e) Surprise - AU R2

### 3.3.2 Region of Interest

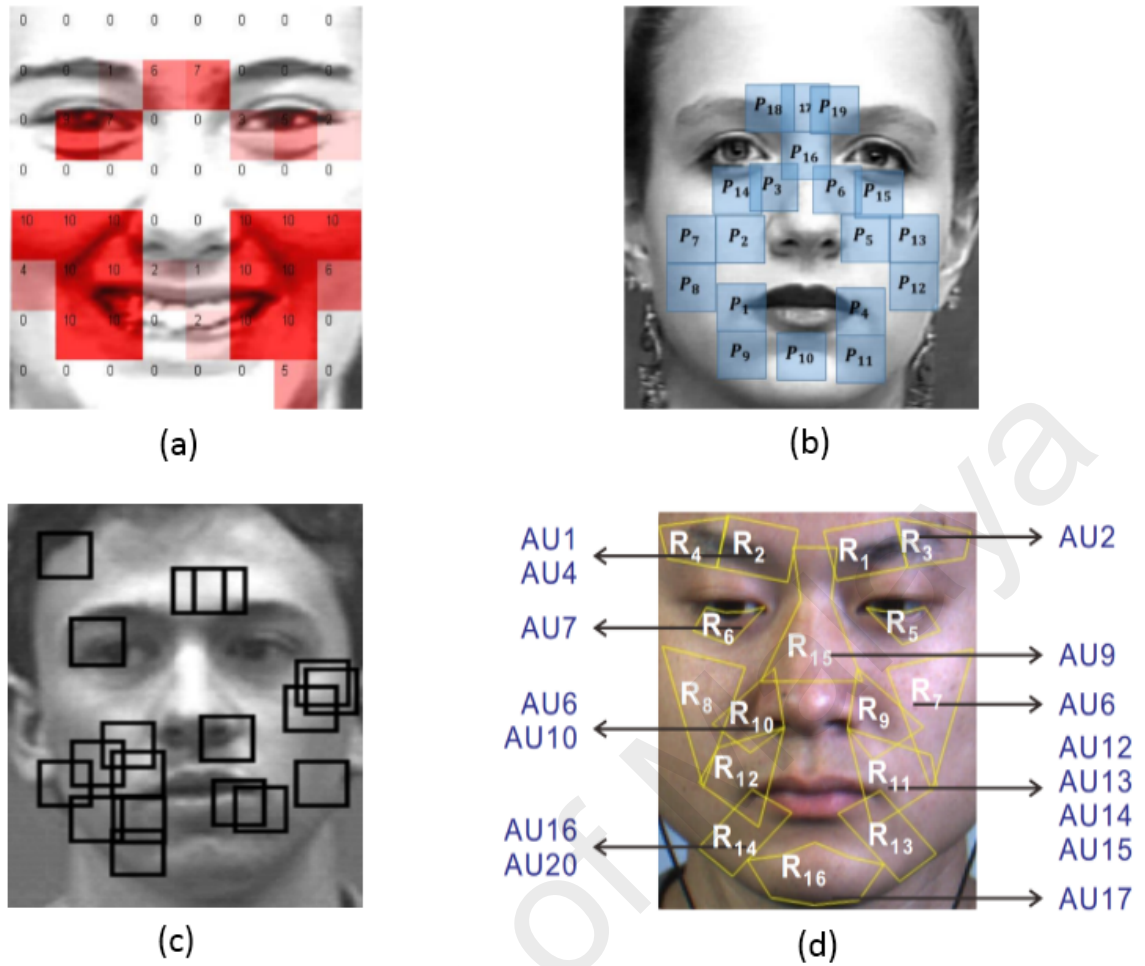
Some recent works on expression analysis achieved high recognition rate by considering a smaller set of descriptive facial patches. For example, Fan and Verma (Fan & Verma, 2009) employed a fusion of four facial regions (i.e., left eye, right eye, nose and mouth) that contain the most discriminative facial characteristics on human faces to perform face recognition on FERET database (Phillips et al., 1998). They claimed that the classification accuracy obtained using their approach is the best among the previously published results.

Zhong et al. (Zhong et al., 2015) proposed a two-stage multitask sparse learning framework to extract discriminative patches (total of approximately one-third of the face), mostly located around the mouth, nose and eyes regions, as illustrated in Figure 3.4(a).

They first select some common dominant facial patches across all the expressions, followed by expression-specific facial patches learned with the aid of face verification. On a similar note, another recent method (Happy & Routray, 2015) also extracted salient facial patches by identifying various facial landmarks initialized by coarse RoIs. The example of the selected regions is shown in Figure 3.4(b). Albeit good performance on the macro-expression databases by CK+ (Lucey et al., 2010) and Jaffe (Lyons et al., 1999), their method appears to ignore the importance of temporal dynamics.

An earlier work by (Anderson & McOwan, 2006) utilized motion velocity information by averaging them over specific pre-defined regions (demonstrated in Figure 3.4(c)) of the face in condensed form. The motion values are determined by a robust differential based optical flow algorithm. In detail, these works tailor their patch selection methods and subsequent feature representations towards macro-expressions. As micro-expressions are minute occurrences and can easily be misinterpreted as noises, a learning-free approach coupled with good facial registration are vital factors towards careful region selection.

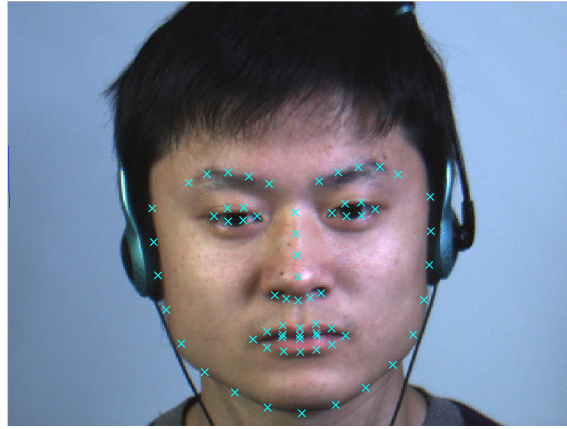
A recent attempt by Wang et al. (S. J. Wang, Yan, et al., 2014) that suggests to extract the micro-expression features from a specific set of facial regions (i.e., 16 RoIs) with fix sizes and shapes, achieves some promising expression recognition results. However, the 16 RoIs are highly dependent on precise estimation of a large number of landmarks and hence it is vulnerable towards registration errors. Moreover, the regions are locked to a fixed size shape dictated by the landmarks, which may not be the optimal areas that capture the perfect feature information all the time. Figure 3.4(d) shows that the regions are determined based on AU knowledge.



**Figure 3.4:** Regions of interest suggested by: (a) Zhong et al. (Zhong et al., 2015); (b) Happy and Routray (Happy & Routray, 2015); (c) Anderson and McOwan (Anderson & McOwan, 2006); (d) Wang et al. (S. J. Wang, Yan, et al., 2014)

### 3.3.3 Landmark Coordinate Detector

To extract the features of the facial RoIs at particular locations, it is essential to register and track the facial feature points in the pre-processing stage. As mentioned earlier in Chapter 2.2.1, there are three techniques to detect the sets of feature points: (a) ASM (Van Ginneken et al., 2002); (b) CLM (Cristinacce & Cootes, 2006), and; (c) AAM (Cootes et al., 1998). Among the feature point detectors, CLM has been reported to be more robust, effective and precise in tracking when compared to the holistic-based model (i.e., AAM) method, as demonstrated by Cristinacce and Cootes (Cristinacce & Cootes, 2008). They examined the superiority of CLM over AAM by conducting several sets of experiment on medical images, such as magnetic resonance brain images, dental



**Figure 3.5:** Example of annotating the 66 landmark coordinates using DRMF method on a CASME II image

panoramic tomograms and human faces. In general, CLM learns the shape model and the variation in appearance from a labeled training set and generates a set of template regions surrounding each individual feature point (Cristinacce & Cootes, 2004).

In 2013, Asthana et al. (Asthana et al., 2013) proposed a fully automatic and relatively quick facial landmark detector – DRMF. It is constructed based on the CLM framework. Using the MATLAB implementation by the authors, it only takes one second of execution time per image (Asthana, n.d.). In contrast to AAM, it adopts part-based approach, where each image patch around the landmark points captures the local texture properties of the object. The sampled regions are then projected onto a reference frame and an efficient shape constrained search using normalized correlation will generate a set of response surface maps. It has been reported to outperform previous landmark detection methods (Zhu & Ramanan, 2012; Saragih et al., 2011), with lower computational time, with features including real-time capabilities and ability to handle faces in-the-wild. In addition, this framework is the current state-of-the art among all fitting optimization strategies for CLM.

An example of annotating the 66 landmark coordinates using DRMF method on an image from CASME II is shown in Figure 3.5.

### 3.3.4 Feature Representation

Besides locating the most salient areas, facial information should also be well-represented in order to properly characterize expressions. For the purpose of experiments of this chapter, LBP-TOP and optical strain are utilized to describe the facial features. Brief information of these two feature descriptors are highlighted below.

As mentioned in Chapter 2.3.2, LBP-TOP has been extensively researched and many variants exist in the literature, with a majority of contribution coming from the area of texture classification (Guo et al., 2010) and facial analysis (Shan et al., 2009). There are many benefits of employing LBP-TOP as feature descriptor, such as high discriminating power, computational simplicity, capability in capturing spatio-temporal detail, concise texture representation as well as robustness to rotation, translation, and illumination change. Recently, LBP-TOP descriptor has also found its way to micro-expression recognition (Yan, Li, et al., 2014; Y. Wang et al., 2014; Li et al., 2013).

On the other hand, recall from Chapter 2.3.4, Shreve et al. (Shreve et al., 2009) verified that expressing the tiny facial movement information using optical strain patterns resulted in better performances when distinguishing small motion of a two-dimensional deformable object in an image. They reported a perfect 100% micro-expressions detection accuracy on their own USF database. In their later work (Shreve et al., 2014), by validating a newer approach on the USF-Combination dataset with more samples, the optical strain technique achieved around 80% true positive rate for detecting the micro-expression frames.

### 3.3.5 Databases

Two most recent and comprehensive databases that meet the requirements of spontaneous micro-expression are considered in this chapter, namely SMIC-HS and CASME II. The detailed information of SMIC-HS and CASME II datasets have been elaborated in Chap-



ter 2.4.2.1 and Chapter 2.4.4.1, respectively. In brief, they are both publicly available and comprise of sufficiently large number of video samples for experimental evaluation. Samples from both SMIC-HS and CASME II are acquired at relatively high frame rates (i.e., more than 100 fps) to better locate the occurrence of micro-expressions. Hence, they are conducive for micro-expression recognition research. Since micro-expression recognition research is still at an early stage, both databases are recorded in a constrained laboratory condition with the labeling done by two trained annotators. Unemotional facial movements are also eliminated from the final selected sequences to allow more relevant expression detail to be captured in the feature extraction process.

### **3.4 Proposed Facial Regions Selection**

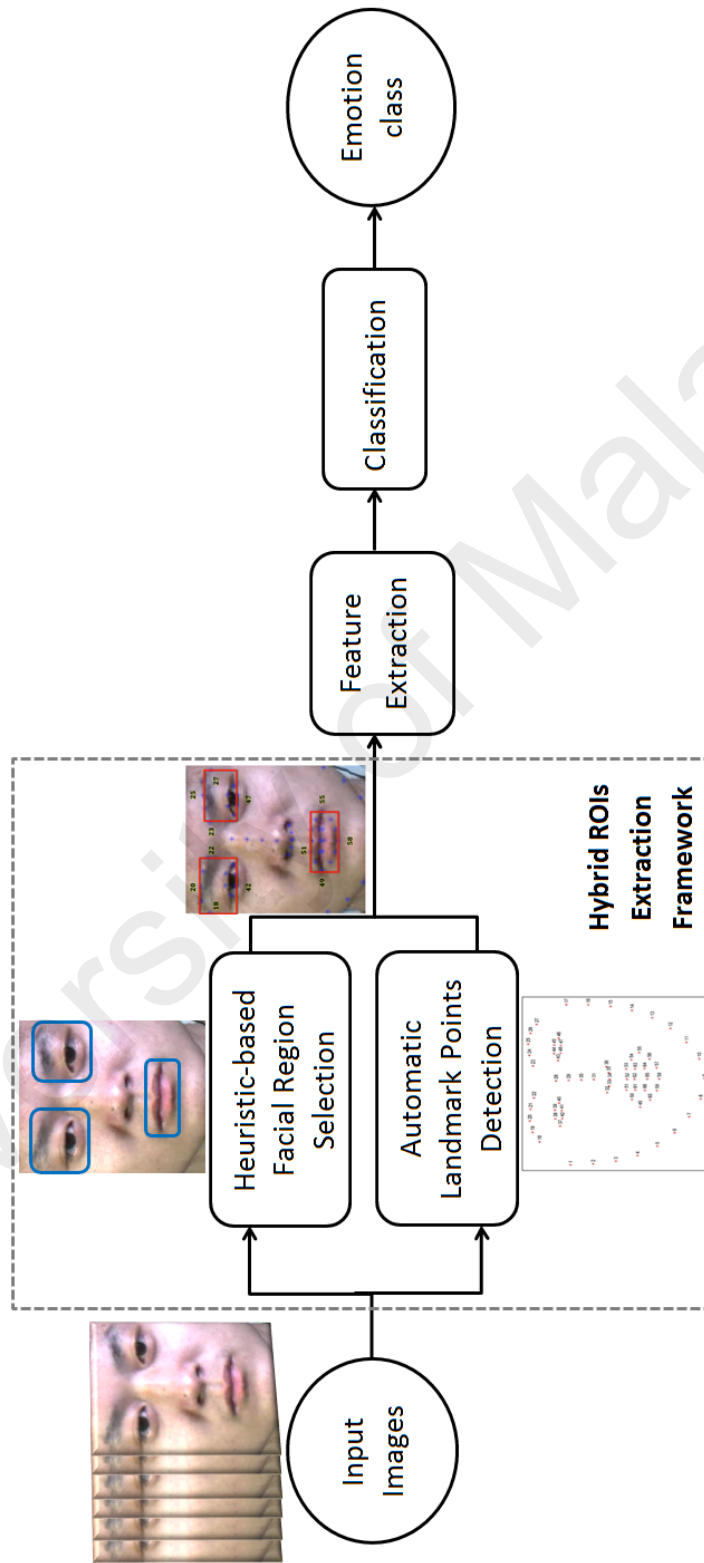
In this chapter, a novel hybrid RoIs selection approach is proposed, namely *RoI-Selective*, to better recognize spontaneous micro-expressions. To achieve this goal, the RoIs selector and two feature extractors are put forward as follows:

1. Hybrid RoIs extraction - the position of 66 facial landmark points are automatically annotated by using the DRMF method. Then, the three RoIs determined heuristically from the database which contain significant and important micro-expression information are extracted from the entire frame, based on the coordinate of selected landmarks.
2. OSF feature extraction - Optical Strain Features (OSF) is employed to describe the features to compare the reliability of the expression information extracted from the RoIs only to that of the entire face. For each RoI in each frame, the optical strain magnitudes in each pixel within the RoIs are calculated from the approximated optical flow values. Next, the strain magnitudes of the three RoIs are concatenated into a single row vector and directly represented as features.

3. LBP-TOP feature extraction - to further verify the effectiveness of the RoIs-based approach, block-based LBP-TOP is utilized as the second feature descriptor. Similar to OSF, the features are extracted strictly from those three RoIs only.

The flowchart of the proposed algorithm is illustrated in Figure 3.6, where the feature extraction process is either OSF or block-based LBP-TOP.

University of Malaya



**Table 3.1:** Emotion description in terms of facial action units

Emotions	Criteria
Happiness	AU 6 or AU 12
Disgust	AU 9 or AU 10 or AU 4+7
Surprise	AU 1+2, AU 25 or AU 2
Repression	AU 15 or AU 17 alone or in combination
Tense	AU 4 or AU 14 or AU 17

**Table 3.2:** Frequency of the face regions based on the action units for five emotions

Face Regions	AU(s) in the region	Emotions	Frequency
Eye + Eyebrow	1, 2, 4, 7	Tense, Disgust, Surprise	5
Mouth	10, 12, 14, 15, 25	Happiness, Repression, Tense, Disgust, Surprise	5
Chin	17	Repression, Tense	2
Cheek	6	Happiness	1
Nose	9	Disgust	1

### 3.4.1 Hybrid RoIs Extraction Approach

In the RoIs extraction fusion process, the micro-expression details are enriched by combining the information from the observed data with the detected landmark point features. Firstly, the RoIs are empirically selected according to the frequency of the AUs during the existence of micro-expressions. Table 3.1 shows the AUs corresponding to the emotions provided in the CASME II database (Yan, Li, et al., 2014). Table 3.2 summarizes the occurrence of the face regions based on the AUs. As shown, the highest frequency regions (i.e., “eye + eyebrow” and “mouth”) are chosen as the RoIs because these regions contribute the majority and meaningful micro-expression details among all other areas of the face.

Secondly, to compare the viability and superiority of the features extracted from the selected RoIs as opposed to the entire image, the RoIs are extracted out from the face. In order to automatically remove the unwanted areas of the image, all 66 landmark points from the first frame of each video are detected by DRMF automated landmark detector (Asthana et al., 2013). In the DRMF method, the tree-based face detector is used to

**Table 3.3:** The landmark points determining the corresponding RoIs bounding boxes

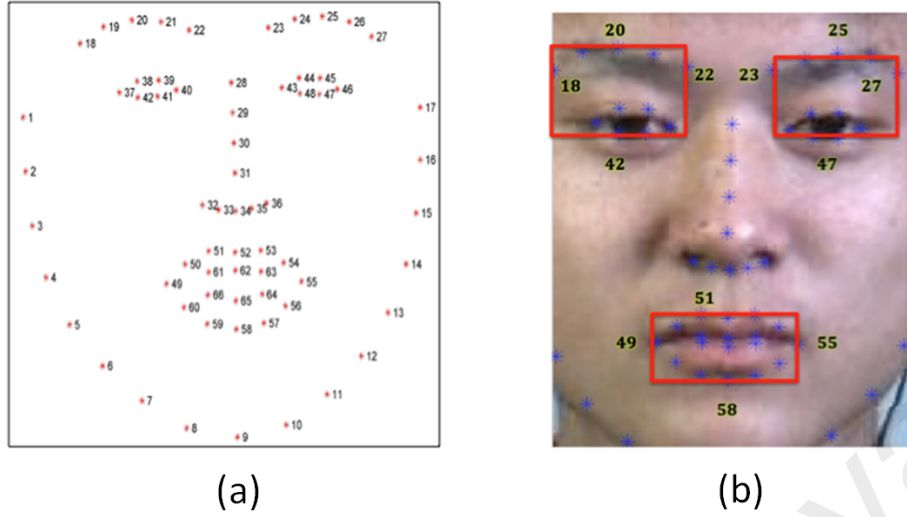
RoIs	Top	Bottom	Left	Right
Left eye + left eyebrow	19 or 20 or 21	41 or 42	18	22
Right eye + right eyebrow	24 or 25 or 26	47 or 48	23	27
Mouth	51 or 52 or 53	58	49	55

achieve high accuracy landmark detection. This landmark annotation step is only applied once (i.e., the first frame only) in each video. The reason is that the motion of the subject is very small and the video frame rate is high (i.e., more than 100 fps). Therefore, the landmark points for the remaining frames are assumed to be similar to those in the first frame.

Next, the bounding boxes of the RoIs are determined according to the neighboring landmark points. All three RoIs (i.e., “left eye + left eyebrow”, “right eye + right eyebrow” and “mouth”) are extracted in multiple rectangular boxes. Precisely, the designated landmark points for each RoI are shown in Table 3.3. The size of each rectangular box depends on the four borders (i.e., top, bottom, left and right) of the corresponding landmark points. Hence, all RoIs have different dimensions in different videos.

There are three benefits to consider only the three RoIs for feature representation instead of the entire facial image: (a) to discard the unnecessary parts of the face that do not contain any facial emotions; (b) to eliminate the existing background noises captured by the camera, which may affect the original pixel intensities, and; (c) to reduce the computation time in feature extraction process (i.e., forming the histogram using OSF and LBP-TOP) due to smaller input size.

The process to segment out the three RoIs from an image with the aid of the DRMF tool is illustrated in Figure 3.7.



**Figure 3.7:** Cropping out the three RoIs: (a) The 66 landmark points marked by DRMF; (b) The rectangular boxes are set based on the coordinate of the 12 landmark points of the four borders

### 3.4.2 Optical Strain Features (OSF) feature extractor

Optical strain is capable of capturing tiny motion between two adjacent frames because each optical strain magnitude contributes to the information of the expression at the pixel level. The feature extractor, Optical Strain Feature (OSF) is built mainly from optical strain.

For clarity, the notations used in this chapter are first explained in detail. A micro-expression video clip is expressed as:

$$s_i = \{f_{i,j} | i = 1, \dots, n; j = 1, \dots, F_i\}, \quad (3.1)$$

where  $s_i$  is the number of videos in the database.  $F_i$  is the total number of frames in the  $i$ -th video sequence, which is taken from a collection of  $n$  video sequences.

The magnitude of the optical flow in each position  $(x, y)$  is computed by estimating the motion between two frames. Then, the optical strain magnitudes for each position  $(x, y)$ , denoted by  $\varepsilon_{x,y}$  (obtained from Equation (2.16)), is calculated over each flow field for two adjacent frames  $(f_{i,j}, f_{i,j+1})$  in the video. Hence, each video contains  $F_i - 1$  number of optical strain maps.

Each video of resolution  $X \times Y$  produces a set of  $F_i - 1$  optical strain maps,  $m_{i,j}$ , each denoted by:

$$m_{i,j} = \{\epsilon_{x,y} | x = 1, \dots, X; y = 1, \dots, Y\}, \quad (3.2)$$

for  $i \in 1, \dots, n$  and  $j \in 1, \dots, F_i - 1$ .

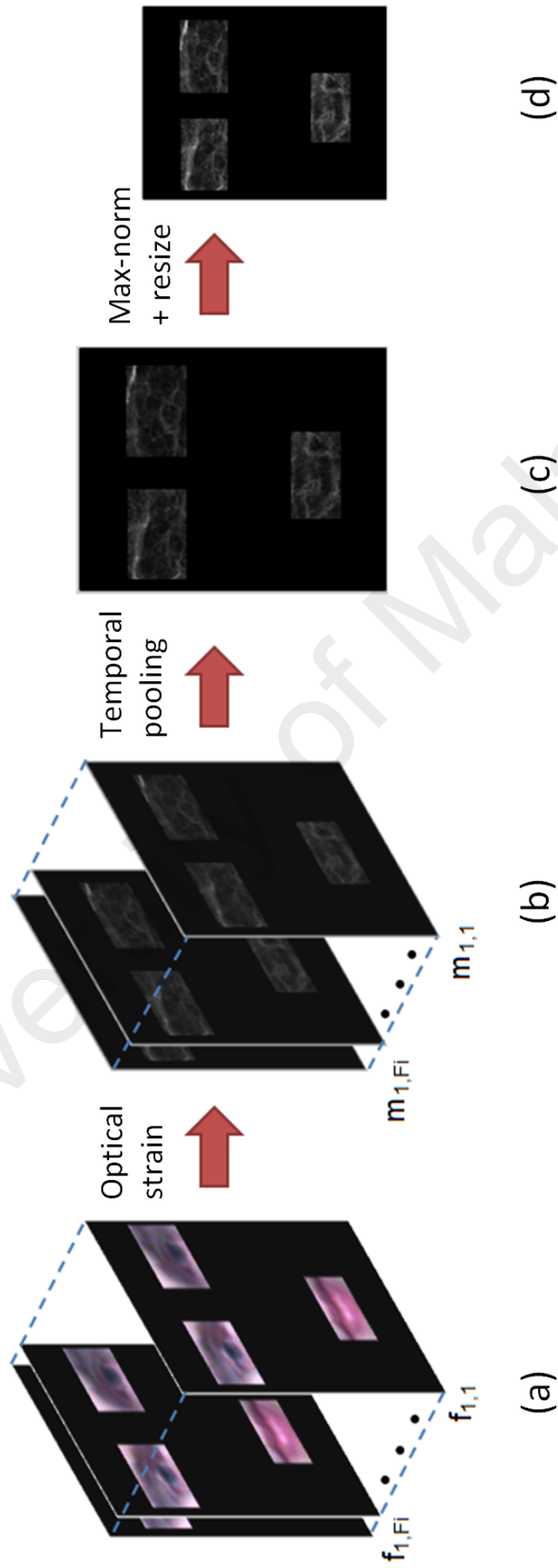
Since only three RoIs are considered,  $\epsilon_{x,y}$  within the RoIs are computed along all strain maps in each video  $s_i$ . To summarize the strain values over time into a compact representation, temporal pooling is performed by summing up  $\epsilon_{x,y}$  in each pixel location of all the strain maps in each  $m_i$  video sequence. The resultant values of  $\epsilon_{x,y}$  are then divided by  $F_i - 1$  because each video has different number of frames. Specifically, to standardize the range of  $\epsilon_{x,y}$  as well as improving their significance, maximum normalization is performed. It is because the feature length of each video needs to be equal before proceeding to the classification stage. Therefore, all the pooled strain RoIs are fixed to constant resolution of  $r \times r$  pixels by using bilinear interpolation. The process flow of extracting the OSF from the three RoIs is shown in Figure 3.8.

Figure 3.9 shows the comparison of the normalized optical strain magnitude between the ‘entire face’ and the three ‘RoIs’, to differentiate the optical strain magnitudes captured by them. It is observed that the proposed ‘RoIs’ approach yields a distinct peak optical strain magnitude, while the ‘entire face’ approach may yield more than one peak (e.g., see Figure 3.9(b)). In addition, for video in which both approaches each yields a distinct peak, the frame distance from the peak frame to the apex frame is smaller in the case of ‘RoIs’ (i.e., see Figure 3.9(a),(c)). This implies that the utilization of three RoIs can better distinguish the frame index that has larger motion (i.e., apex frame). In contrast, there may also be some ambiguity in locating a distinct peak frame (Figure 3.9(b)) when the entire face is considered. Therefore, the utilization of three ROIs is indeed ben-

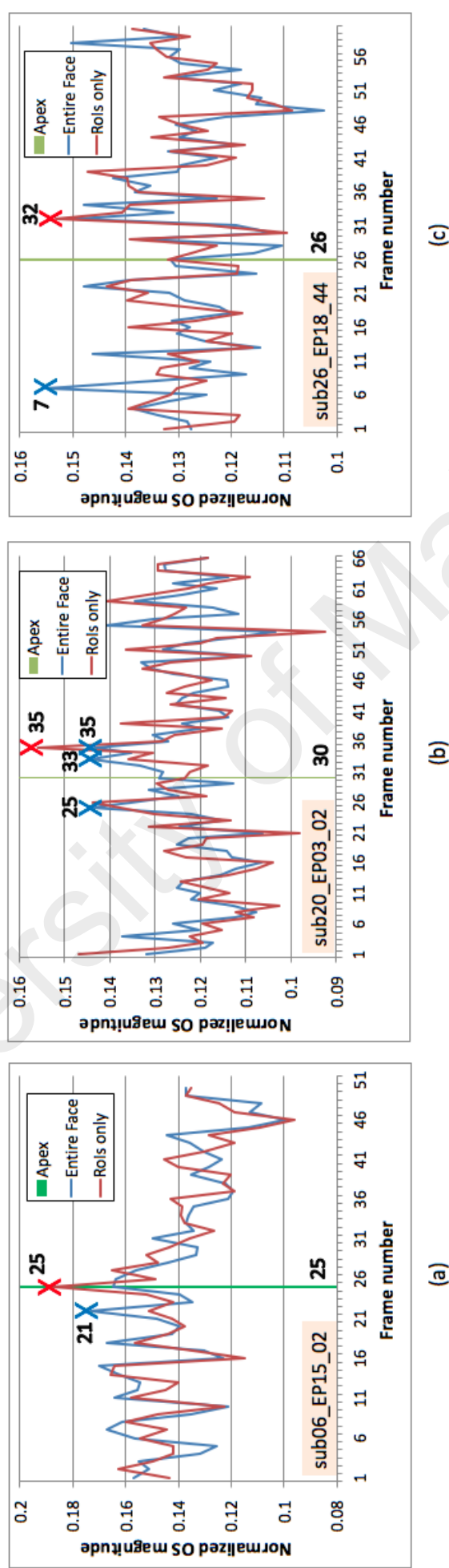
eficial for better characterization of local movements rather than holistic utilization of the entire face region.

University of Malaya





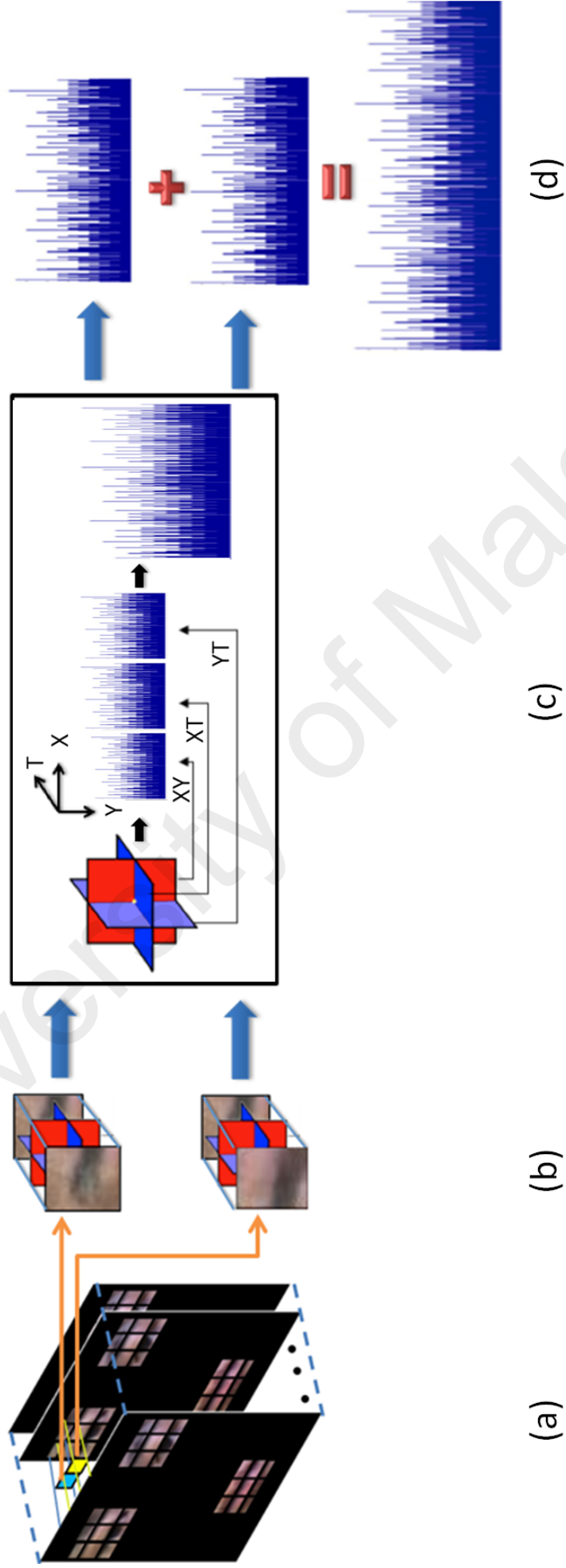
**Figure 3.8:** Optical strain feature extraction for the first video after cropping out the RoIs: (a) Original frames,  $f_{1,j}$ ; (b) Strain maps,  $m_{1,j}$ ; (c) Temporal pooled strain map; (d) Maximum-normalized and resized frame



**Figure 3.9:** Comparison of the normalized optical strain magnitude between the three RoIs (taken together) and the entire face region along a sample video sequence

### 3.4.3 Block-based LBP-TOP feature extractor

The second feature descriptor considered is block-based LBP-TOP, which represents the features in the spatio-temporal perspective. The three RoIs from the frames are first segmented into  $N \times N$  non-overlapping blocks in order to obtain the features that are local to each block region. Next, the features extracted from the three orthogonal planes (i.e.,  $XY$ ,  $XT$  and  $YT$ ) are concatenated to form the final histogram. Refer to Chapter 2.3.2 for detailed descriptions of block-based LBP-TOP. The procedure to construct the LBP-TOP features from the three RoIs is illustrated in Figure 3.10.



**Figure 3.10:** Features of the first two blocks volumes extracted by using block-based LBP-TOP: (a) Each RoI is partitioned into  $3 \times 3$  blocks; (b) LBP features generated from  $XY$ ,  $XT$  and  $YT$  planes; (c) Concatenation of features in each block into a single histogram; (d) Concatenation of block histograms to form the final histogram

### 3.5 Experiments

#### 3.5.1 Datasets

Based on the pros and cons of the micro-expression databases discussed in Chapter 2.4, CASME II (Yan, Li, et al., 2014) and SMIC-HS (Li et al., 2013) datasets are selected to evaluate the proposed algorithm with the primary benefits of large sample size and high frame rate. Brief information of these datasets is described as follows.

The CASME II dataset consists of 26 subjects and a total of 246 sequences of micro-expressions, with one video per sequence. It contains five classes of emotions, namely, happiness (32 samples), disgust (60 samples), surprise (25 samples), repression (27 samples) and tense (102 samples). The baseline performance reported is 63.41% using  $5 \times 5$  non-overlapping block-based LBP-TOP as the feature extractor and SVM classifier with LOVOCV configuration as the classifier.

On the other hand, the SMIC-HS dataset consists of 164 micro-expression samples from 16 participants. There are three main emotion categories: positive (happiness; 51 samples), negative (sad, fear, disgust; 70 samples), and surprise (43 samples). The three-class baseline performance reported is 48.78%. Block-based LBP-TOP with  $8 \times 8$  partition blocks and SVM with LOSOCV configuration are employed as feature descriptor and classifier, respectively.

#### 3.5.2 Experiment Settings

The experiments are conducted on the CASME II and SMIC-HS datasets by employing the feature descriptors of OSF and block-based LBP-TOP. SVM classifier with a polynomial kernel of degree 6 is used to examine the proposed *RoI-Selective* approach. In this multi-subject level analysis, both LOSOCV and LOVOCV are utilized to validate the effectiveness of the proposed approach in all the experiments.

Due to the imbalanced distribution of expression class samples in the databases,  $F$ -



**Figure 3.11:** A sample video sequence of ‘Surprise’ micro-expression from SMIC-HS dataset

*measure*, *recall* and *precision* metrics are also adopted to evaluate the performance of the proposed approach, as suggested by Le Ngo et al. (Le Ngo et al., 2014). This supplements the conventional accuracy rate, which may not provide a sufficiently fair gauge. Specifically, recall (exactness) is the ratio of the relevant information extracted by the system to the total number of relevant records in the database, and precision (completeness) is the measure of how much information in the system is returned correctly. F-measure is the harmonic mean of precision and recall. The equations of these three indicators are set out as follows:

$$\text{Recall} := \frac{TP}{TP + FN}, \quad (3.3)$$

$$\text{Precision} := \frac{TP}{TP + FP}, \quad (3.4)$$

$$\text{F-measure} := 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (3.5)$$

where TP, FN and FP are true positive, false negative and false positive, respectively.

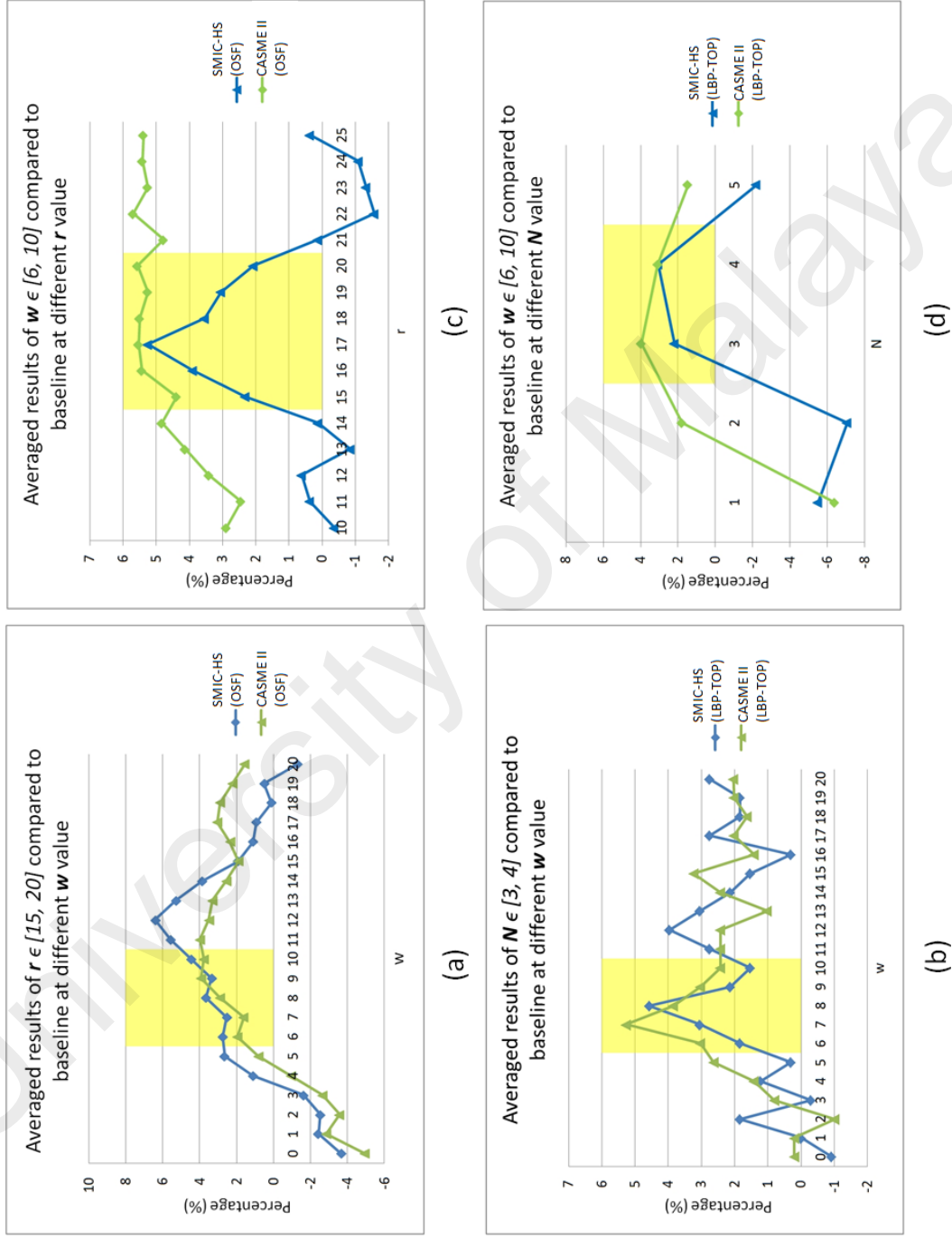
For extensive experiment purpose, augmentation of the RoI window size (for all three RoIs) is performed by enlarging the RoI window by  $w$  pixels towards all four directions (top, bottom, left and right). This enlargement of area further enriches the expression information in each RoI, and provides another free parameter for further experimentation. The best performances in this experiment can be attained by enlarging the RoI areas by  $w \in [6, 10]$  while the RoI areas are resized to  $r \in [15, 20]$  when the OSF method is utilized for feature extraction. Meanwhile, the LBP-TOP features in each RoI are extracted from  $N \times N$  non-overlapping blocks, with  $N \in [3, 4]$ .

### 3.6 Results and Discussions

In this sub-chapter, the parameters to be used are first analyzed and discussed, followed by the best recognition performances obtained from the optimal range of values and the computational cost of the system.

#### 3.6.1 Parameter Analysis

The three free parameters, namely,  $w$ ,  $r$  and  $N$ , provide us with further clues on the importance of using the three essential RoIs for processing. More precisely, the OSF and LBP-TOP methods both have two parameters each, namely  $r$  and  $w$  for OSF,  $N$  and  $w$  for LBP-TOP (see Figure 3.12); hence only two parameters are tuned for each feature extraction method. Instead of increasing or reducing these parameter values in an ad-hoc manner, it is crucial to analyze and identify the ideal ranges for these parameters systematically. For instance, in the case of OSF, the value of one particular parameter is set (i.e.,  $w = 6$ ) and the average result of the other parameter is determined within a certain range (i.e., average result of  $r \in [15, 20]$ ). Figure 3.12 reports an analysis of the results obtained for all four pairings of parameters.



**Figure 3.12:** Results (Percentage of improvement in accuracy over baseline) of various combination of parameter settings by holding the value of: (a)  $w$  in OSF,  $r \in [15, 20]$ ; (b)  $w$  in LBP-TOP,  $N \in [3, 4]$ ; (c)  $r$  in OSF,  $w \in [6, 10]$ , and; (d)  $N$  in LBP-TOP,  $w \in [6, 10]$



These parameters are not only intuitive but also provide meaningful insights into the spatial information found within the frames that contain micro-expression. The three parameters, i.e., locality ( $w$ ), dimensionality ( $r$ ) and scale ( $N$ ) are described in detail as follows:

- **Locality:**  $w$  tunes the spread or “locality” of useful features within each RoI. Larger  $w$  might include extra regions that do not contribute to subtle movements while smaller  $w$  might result in missing crucial information around the RoI. This trend can be clearly seen in Figure 3.12(a) and (b). Interestingly, the RoI-based approach performs worse than the baselines for both databases when  $w$  is very low (i.e.,  $0 \sim 3$ ). On the other hand, high  $w$  values are also bad choices, likely due to encroachment of the regions into the nose area or area above eyebrows.
- **Dimensionality:**  $r$  adjusts the resolution or to be more precise, the “dimensionality” of the OSF-extracted features. This measures how much information from the OSF are encoded. Larger  $r$  encodes more information, and vice versa. In Figure 3.12(c), performance on CASME II improves when the dimension of OSF increases. However, the performance on SMIC-HS is unusual, probably due to the fact that SMIC-HS has a smaller average sample area size (difference of averages is approximately  $25 \times 25$  pixel), hence more sensitive towards the resizing of the RoIs. An optimum range is more desirable for SMIC-HS.
- **Scale:**  $N$  is a typical parameter from the block-based LBP-TOP algorithm. Intuitively, it “scales” the features to be extracted into histograms. Larger  $N$  describes the statistics of features more locally, while smaller  $N$  will provide more global statistics for the RoIs. Considering that the RoIs themselves are already much smaller in size compared to the original frame, overly sub-divided blocks (e.g.,  $N = 5$ ) might negatively affect the intrinsically captured patterns, while using a sin-

gle histogram (i.e.,  $N = 1$ ) misses the localized patterns and is equally detrimental to the results.

From the plots, the range of parameters that yield consistently good results is highlighted in yellow. In particular,  $w \in [6, 10]$  is best suited for both OSF and LBP-TOP as suggested by Figure 3.12 (a) and Figure 3.12(b) ;  $r \in [15, 20]$  for OSF as suggested by Figure 3.12(c), and;  $N \in [3, 4]$  for LBP-TOP as suggested by Figure 3.12(d). These are also the optimal ranges of operation for the parameters that will be reported in Chapter 3.6.2.

### 3.6.2 Recognition Performance

The baseline performance is established by reproducing results from the original SMIC (Li et al., 2013) and CASME II (Yan, Li, et al., 2014) works. The baseline methods consider the entire facial region for block-based LBP-TOP (Zhao & Pietikainen, 2007) feature extraction. In addition, optical strain features (OSF) (Liong, Phan, et al., 2014) is also applied as an alternative feature for more comprehensive comparisons. Following the original implementations,  $LBP-TOP_{4,4,4,1,1,3}$  with  $8 \times 8$  blocks for SMIC and  $LBP-TOP_{4,4,4,1,1,4}$  with  $5 \times 5$  blocks for CASME II are applied. The baseline results are summarized in Table 3.4, with the accuracy measure used in accordance to the original works.

For the proposed ROI selection scheme, the optimal range of parameters (i.e.  $r$ ,  $w$  and  $N$ ) used are described in Section 3.6.1. As demonstrated in Figure 3.13, the utilization of these selected facial RoIs results in promising performance on both micro-expression databases considered in this study. The best parameter values from Figure 3.6.1 for different cases yield a performance improvement of approximately:

- 10% at  $w = 10$  and  $r = 17$  with OSF in LOVOCV
- 5.5% at  $w = 8$  and  $N = 4$  with LBP-TOP in LOVOCV

**Table 3.4:** Reproduced baseline recognition results (%)

	LOSOCV	LOVOCV
SMIC-OSF	38.70	46.34
SMIC-LBPTOP	48.17	59.15
CASME-OSF	26.72	48.99
CASME-LBPTOP	35.22	58.30

- 8% at  $w = 10$  and  $r = 16$  with OSF in LOSOCV
- 10.5% at  $w = 7$  and  $N = 4$  with LBP-TOP in LOSOCV

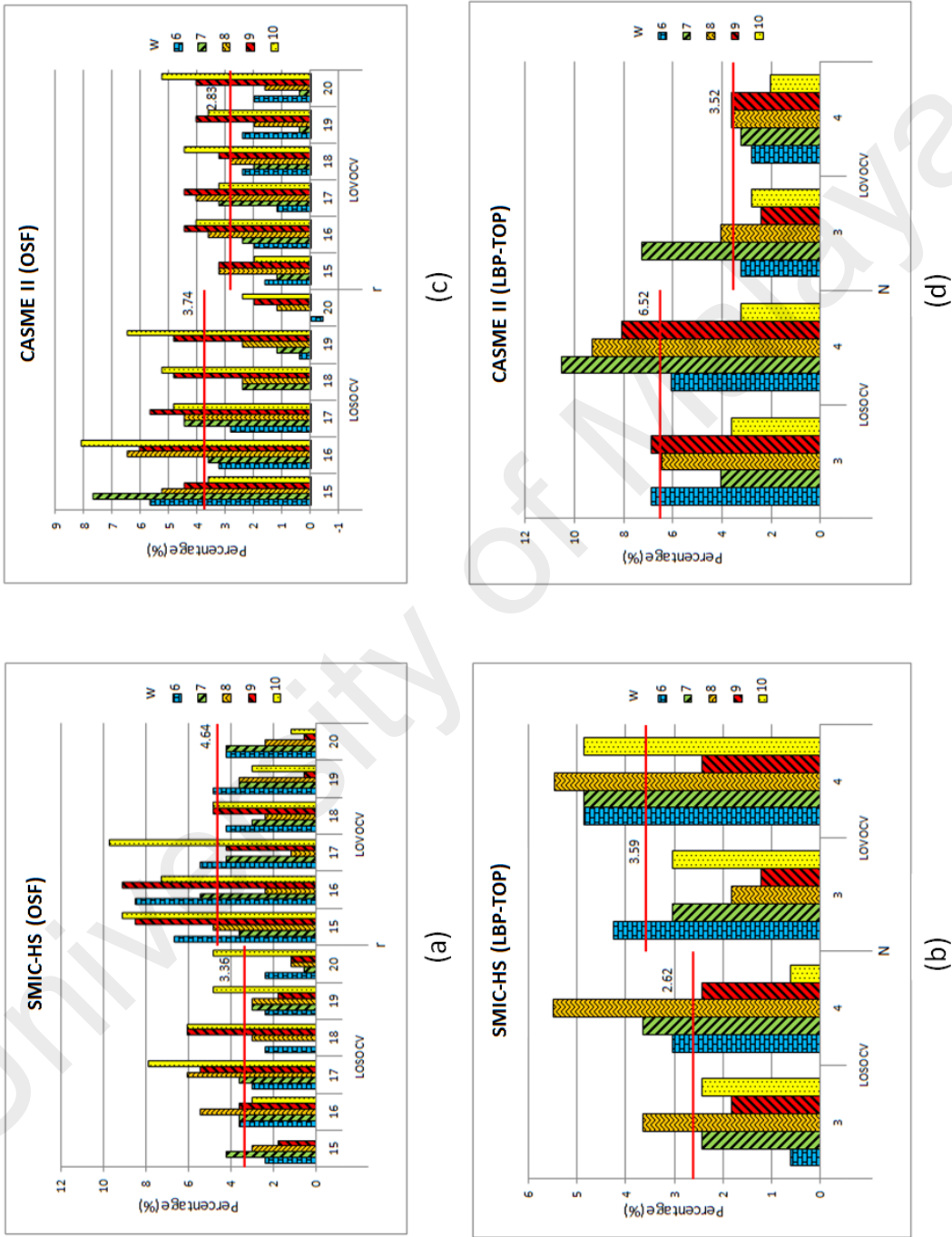
At a more realistic level, the average increment for each scenario is also provided (see red horizontal line in each plot in Figure 3.13) to gain better insight into the actual averaged performance for each case. Overall, the RoI-based approach produces an average improvement of around  $2.6 \sim 4.6\%$  for the SMIC-HS dataset, and around  $2.8 \sim 6.5\%$  for the CASME II database. There is a common trend observed among the plots, that is, SMIC-HS always performs better under LOVOCV settings while CASME II has a greater improvement when the LOSOCV method is considered. It suggests that the data fusion strategy on selecting the salient facial regions has significant improvement on classifying the micro-expressions compared to considering the entire facial region.

Table 3.5 shows the F-measure, precision and recall results obtained by the proposed *RoI-selective* approach against their respective baselines. In all cases, the proposed method outperforms the baseline results, except for the solitary case of LOSOCV in CASME II using the OSF feature extractor. However, in general, the proposed approach is capable of yielding consistently good results for micro-expression recognition when compared to the full-face image.

On top of that, the recognition performance of the *RoI-Selective* approach is compared to the existing methods with the measurements of Accuracy, F-measure, recall and precision score, as tabulated in Table 3.6. It can be seen that the *RoI-Selective* approach is

capable of outperforming the other methods significantly. Notably, substantially superior results are obtained in SMIC-HS database.

University of Malaya



**Figure 3.13:** The percentage of improvement in recognition accuracy achieved by varying parameters  $w$  with  $r$  or  $N$  using: (a) OSF method for SMIC-HS; (b) LBP-TOP method for SMIC-HS; (c) OSF method for CASME II; (d) LBP-TOP method for CASME II

### 3.6.3 Discussion on Computational Cost

In addition, execution time is evaluated by randomly selecting one of the videos in SMIC-HS with LBP-TOP as the feature descriptor. Running on an Intel Core i7-4770 CPU 3.40 GHz machine, the time taken for the feature extraction process by the baseline method (Li et al., 2013) is 0.0775 s per frame while it takes 0.0642 s per frame in the proposed *RoI-Selective* approach. Based on this result, *RoI-Selective* approach achieves a speed-up of approximately 17% compared to the baseline method due to the savings from the smaller input size. Moreover, the feature dimension of the baseline method is 2,880 per video sample while it is only 1,215 in the proposed *RoI-Selective* approach. With a much smaller set of features (by nearly 58% of the original set), feature complexity at the classification stage can also be reduced, which in turn decreases the overall computation time.

**Table 3.5:** F-measure, recall and precision scores of the proposed *RoI-selective* approach against their respective baselines in four different scenarios (averaging across the good parameter ranges in Figure 3.13)

Methods	LOSOCV			LOVOCV		
	F-measure	Recall	Precision	F-measure	Recall	Precision
<i>(a) SMIC-HS (OSF)</i>						
Baseline (Liong, Phan, et al., 2014)	.38	.38	.37	.44	.43	.44
<i>RoI-Selective</i>	<b>.41</b>	<b>.41</b>	<b>.41</b>	<b>.50</b>	<b>.49</b>	<b>.51</b>
<i>(b) SMIC-HS (LBP-TOP)</i>						
Baseline (Li et al., 2013)	.49	.50	.48	.59	.56	.63
<i>RoI-Selective</i>	<b>.52</b>	<b>.52</b>	<b>.51</b>	<b>.63</b>	<b>.61</b>	<b>.64</b>
<i>(c) CASME II (OSF)</i>						
Baseline (Liong, Phan, et al., 2014)	.23	.21	.25	.39	.36	.42
<i>RoI-Selective</i>	<b>.22</b>	<b>.22</b>	<b>.22</b>	<b>.45</b>	<b>.44</b>	<b>.46</b>
<i>(d) CASME II (LBP-TOP)</i>						
Baseline (Yan, Li, et al., 2014)	.15	.18	.12	.52	.50	.55
<i>RoI-Selective</i>	<b>.32</b>	<b>.29</b>	<b>.35</b>	<b>.57</b>	<b>.54</b>	<b>.60</b>

**Table 3.6:** Comparison of recognition results of the proposed method to existing methods in measurements of Accuracy, F-measure, recall and precision scores in LOSOCV protocol

#	Methods	SMIC-HS				CASME II			
		Accuracy	F-measure	Recall	Precision	Accuracy	F-measure	Recall	Precision
1	Le et al. (Le Ngo et al., 2014)	.44	.47	.74	.40	.44	.33	.53	.29
2	Wang et al. (Y. Wang et al., 2014)	.38	.39	.40	.38	.46	.38	.32	.47
3	Liong et al. (Liong, See, et al., 2014)	.53	.54	.55	.53	.42	.38	.36	.41
4	Oh et al. (Oh et al., 2015)	.34	.35	.35	.34	.46	.43	.35	.55
5	<i>RoI-Selective</i> (OSF)	.45	.41	.41	.41	.36	.22	.22	.22
6	<i>RoI-Selective</i> (LBP-TOP)	<b>.54</b>	<b>.52</b>	<b>.52</b>	<b>.51</b>	<b>.45</b>	<b>.32</b>	<b>.29</b>	<b>.35</b>



### 3.7 Summary

This chapter proposes a novel hybrid approach, namely *RoI-Selective*, that combines the heuristic and automatic approaches in extracting the important facial regions for micro-expressions recognition. The features are extracted from three desired RoIs (specifically the regions of “left eye + left eyebrow”, “right eye + right eyebrow” and “mouth”) that contain significant and valuable micro-expression information. Selection of the RoIs are statistically determined by the frequency of occurrence of the AUs among all the expressions. The facial landmark points are obtained using the DRMF landmark detector. Then, this information is utilized to crop out the RoIs from each frame automatically. Overall, the proposed *RoI-selective* approach demonstrates promising recognition results on both the spontaneous micro-expression databases, namely, SMIC-HS and CASME II. The best average improvement of performance for the SMIC-HS database is around 4.5% for OSF and 3.5% for LBP-TOP by adopting the LOVOCV protocol. As for the CASME II database, *RoI-selective* approach achieved an average increments of 3.7% for OSF and 6.5% for LBP-TOP by adopting the LOSOCV protocol.

This automated micro-expression recognition system can potentially be deployed in applications such as medical diagnosis, national safety, police interrogation and lie detection. For future works, more attention will be devoted to handle the issues of empirical parameter tuning. Tuning of the parameters (i.e.,  $w$ ,  $r$ ,  $N$ ) towards optimum values and other settings (i.e., degree of the polynomial kernel) in the feature extractors and classifiers warrants further investigation to maximize the performance of the system. In addition, noise filtering schemes can be introduced to minimize the presence of noises resulting from the computation of optical strain.

## CHAPTER 4: FEATURE EXTRACTION BASED ON FACIAL STRAIN

### 4.1 Overview

In this chapter, a novel feature extractor for detecting and recognizing micro-expressions is presented, by utilizing facial optical strain magnitudes to construct optical strain features and optical strain weighted features. As mentioned earlier in Chapter 2.3.2, *detection* refers to determining the presence of micro-expressions on the face without identification of its type, whereas *recognition* goes a step further to distinguish the exact state or type of expression shown on the face. This is computationally essential for the relatively new field of spontaneous micro-expression, where subtle expressions can be technically challenging to pinpoint. As discussed in Chapter 2.3.4, optical strain is an extension of optical flow that is capable of quantifying subtle changes on faces and representing the minute facial motion intensities at the pixel level.

Specifically, the feature histogram of each video sample is designed and constructed using optical strain information, following three main processes:

1. Optical Strain Features (OSF) - all the optical strain images in each video are temporally pooled, then the strain magnitudes of the pooled image are treated as features.
2. Optical Strain Weighted Features (OSW) - optical strain magnitudes are pooled in both spatial and temporal directions to form a weighting matrix. The respective weights of each video are then multiplied with the features from the XY-plane extracted by LBP-TOP.
3. Concatenation the OSF and OSW (OSF + OSW) - the feature histograms from steps (a) and (b) are concatenated to form the final feature histogram.

Experiments are conducted on two spontaneous and high speed micro-expression datasets to verify the performance of the proposed OSF + OSW feature extractor. Both the detection and recognition results suggest that the feature extractor that is developed based on optical strain, can effectively reveal and describe the subtle facial muscle changes.

## **4.2 Literature Review**

### **4.2.1 Optical Strain**

Optical strain patterns justify its superiority over the raw image in face recognition as the computation of the magnitudes is based on biomechanics. It is also robust to lighting conditions, heavy makeup and camouflage (Shreve et al., 2010; Manohar et al., 2007). Optical strain pattern is exploited for spotting facial micro-expressions automatically in several databases, including USF, USF-HD, USF-combination, Canal-9, “found videos” and SMIC micro (Shreve et al., 2009; Shreve, Godavarthy, et al., 2011; Shreve et al., 2014). See Chapter 2.3.4 for the details of the algorithms implemented and the outstanding results obtained when validated on these databases. In short, optical strain has demonstrated its superiority over optical flow by producing more consistent results in automatic micro facial expression spotting task. In this chapter, the strengths of optical strain are leveraged to describe suitable features for detection and recognition tasks.

### **4.2.2 Block-based LBP-TOP**

Block-based LBP-TOP has been discussed in Chapter 2.3.2, hence only basic principles are recapped in this sub-chapter. Block-based method in feature extraction process is widely used in detecting or recognizing micro-expressions, as demonstrated in (Zhao & Pietikainen, 2007; Yan, Li, et al., 2014; Pfister et al., 2011; Li et al., 2013). For LBP-TOP (Zhao & Pietikainen, 2007) texture descriptor, it has been broadly used in human activity classification (Mattioli & Shao, 2009), lip-reading (Zhao et al., 2009), and also facial expression recognition (Shan et al., 2009). It describes the space-time

texture of a video volume, which encodes the local texture pattern by thresholding the center pixel against its neighboring pixels. Block-based LBP-TOP partitions the three orthogonal image planes into  $N \times N$  non-overlapping blocks, where the final histogram is a concatenation of histograms from each block volume. This final histogram represents the appearance, horizontal motion and vertical motion of a video.

#### **4.2.3 Pooling**

The large number of pixels in an image or video can be summarized into a more compact representation of lower dimension. Generally, feature pooling in spatial domain is one commonly employed technique, which partitions the image into several regions, before summing up or averaging the pixel intensities of each region. For instance, spatial pooling is employed together with state-of-the-art feature descriptors such as Scale-invariant Feature Transform (SIFT) (Lowe, 2004) and Histograms of Oriented Gradients (HOG) (Dalal & Triggs, 2005) to enhance their robustness against noise and clutter.

On the other hand, temporal pooling is able to summarize the features over a period of time in a compact and efficient manner. Boureau et al. (Boureau et al., 2010) demonstrated that the performance of the recognition algorithm is attributed to the pooling step of feature extraction. Besides that, Hamel et al. (Hamel et al., 2011) examined the performance of automatic annotating and ranking music radio by different combination of pooling methods (i.e., mean, maximum, minimum and variance). Pooling is also adopted by several researchers to vectorize the feature descriptors in calculating the local or global bag of features (J. Zhang et al., 2007; Sivic & Zisserman, 2003).

#### **4.2.4 Image Filtering**

Gaussian filter is a popular filtering technique that is extensively deployed to process digital images containing facial expressions. It is one of the adaptive filters that removes Gaussian noises in image (Forsyth & Ponce, 2002). For example, to track the AUs of

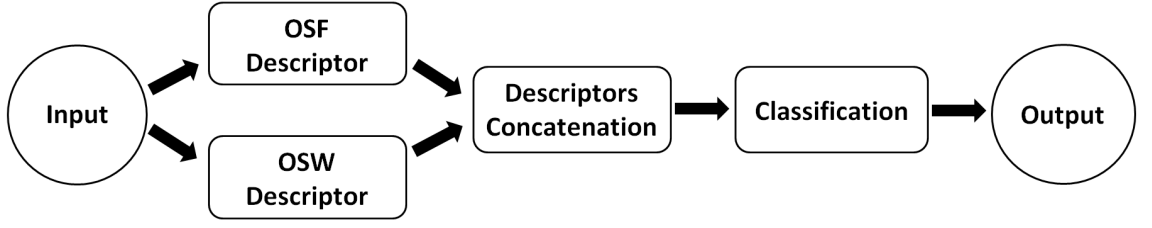
the facial expressions, a 5 x 5 Gaussian filter is applied with different sizes on different regions of the face, to achieve locally smoothing effect (Lien et al., 2000). In facial expression analysis (Z. Liu et al., 2001), Gaussian filter is employed to reduce the noises on the face images in order to compute the illumination change of one person or Expression Ratio Image (ERI) resulting from deformation of the person's face. Detailed advantages and examples of Gaussian filter are provided in Chapter 2.2.2.

### 4.3 Proposed Algorithm

To extract the spatio-temporal features by utilizing facial optical strain information, three main steps are proposed, namely:

1. OSF - the optical strain magnitudes in each frame are derived from the optical flow values. Then all the optical strain maps in each video are temporally pooled into a composite strain map. Thereafter, the optical strain magnitudes in the composite strain map are directly used as the features.
2. OSW - spatio-temporal pooling is applied on the optical strain frames of each video, then the final matrix of normalized coefficient values obtained are used as the weights for each video. The weighting matrix (of  $N \times N$  dimension after pooling) is then multiplied with their respective LBP-TOP-extracted histogram bins on the  $XY$ -plane.
3. OSF + OSW - the feature histograms extracted in steps (a) and (b) are concatenated into a final feature histogram that represents the video sample.

The architecture overview of the flow of the proposed method is illustrated in Figure 4.1.



**Figure 4.1:** Overview of the proposed algorithm

#### 4.3.1 Optical Strain Features

Since optical strain magnitudes can aptly describe the minute extent of facial deformation at the pixel level, they can directly be employed as features as well. This sub-chapter discusses the process of obtaining Optical Strain Features (OSF). The notations used in the subsequent chapters are first described. A micro-expression video clip is expressed as:

$$s_i = \{f_{i,j} | i = 1, \dots, n; j = 1, \dots, F_i\}, \quad (4.1)$$

where  $F_i$  is the total number of frames in the  $i$ -th sequence, which is taken from a collection of  $n$  video sequences.

The optical flow field is first estimated by its 2D motion vector,  $\mathcal{P} = (p, q)$  (from Equation (2.9)). Then, the optical strain magnitude at each pixel location  $\varepsilon_{x,y}$  (from Equation (2.16)) is calculated for each flow field over two consecutive frames, i.e.,  $\{f_{i,j}, f_{i,j+1}\}$ . Hence, each video of resolution  $X \times Y$  produces a set of  $F_i - 1$  optical strain maps, each denoted as follows:

$$m_{i,j} = \{\varepsilon_{x,y} | x = 1, \dots, X; y = 1, \dots, Y\}, \quad (4.2)$$

for  $i \in 1, \dots, n$  and  $j \in 1, \dots, F_i - 1$ .

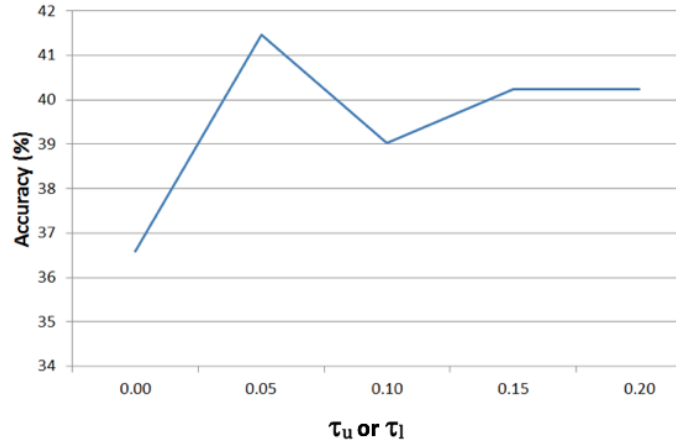
The following steps describe the essential pre-processing steps for noise reduction and signal attenuation, followed by how OSF is obtained.

#### 4.3.1.1 Pre-processing

Prior to feature extraction, two pre-processing steps are carried out to reduce unwanted noises in the optical strain maps.

First, the edges in each strain map  $m_{i,j}$  are removed. Since the edges are the gradient of the moving objects that consist of local maximas, eliminating them is to remove a large number of irrelevant “fake movements” (which will be detected wrongly as facial muscle movements in optical flow estimation later) if the strain map is very noisy (Barcelos et al., 2003). Among the different types of edge detectors, the Sobel filter justifies its feasibility by two main advantages (Gao et al., 2010): (a) its ability to detect the edges in a noisy image by introducing smoothing and blurring effect on the image; (b) the differential of two rows or two columns enhances the strength of important edges. The Sobel operator is a simple approximation to the concept of 2D spatial gradient, by convoluting a grayscale input image with a pair of  $3 \times 3$  convolution mask (Juneja & Sandhu, 2009). Happy and Routray (Happy & Routray, 2015) demonstrated that horizontal edge detector always generates a distinct edge on macro facial expression databases (i.e., CK+ (Lucey et al., 2010) and JAFFE (Lyons et al., 1998)). In this chapter, experiments are conducted on the micro-expression databases to compare the performance of horizontal and vertical edges. It is empirically discovered that removing the vertical edges generate better recognition results than removing the horizontal edges only, as well as both the horizontal and vertical edges. Therefore, Sobel edge detector is employed to spot the vertical directions in this experiment.

Secondly, the magnitudes in each optical strain map  $m_{i,j}$  are clipped to zero for  $\epsilon_{x,y} \notin [\Gamma_l, \Gamma_u]$ , with the two threshold values  $\Gamma_l$  and  $\Gamma_u$  denoting the lower and upper thresholds, respectively. The values of  $\Gamma_l$  and  $\Gamma_u$  are determined using the lower and upper percentages  $(\tau_l, \tau_u)$  of the strain magnitude range, i.e.,  $[\epsilon_{min} = \min\{\epsilon_{x,y}\}, \epsilon_{max} =$



**Figure 4.2:** Effect of  $\tau_l$  and  $\tau_u$  values on micro-expression recognition rate for the SMIC-HS database

$\max\{\varepsilon_{x,y}\}$ . The lower and upper thresholds are computed as follows:

$$\Gamma_l = \varepsilon_{min} + \tau_l \cdot (\varepsilon_{max} - \varepsilon_{min}), \tau_l \in [0, 1], \quad (4.3)$$

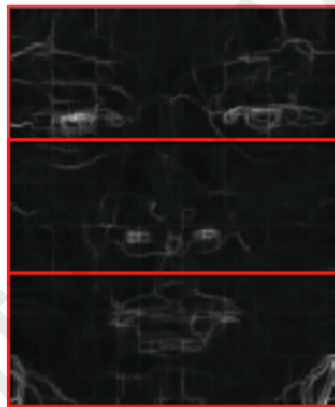
and

$$\Gamma_u = \varepsilon_{max} - \tau_u \cdot (\varepsilon_{max} - \varepsilon_{min}), \tau_u \in [0, 1]. \quad (4.4)$$

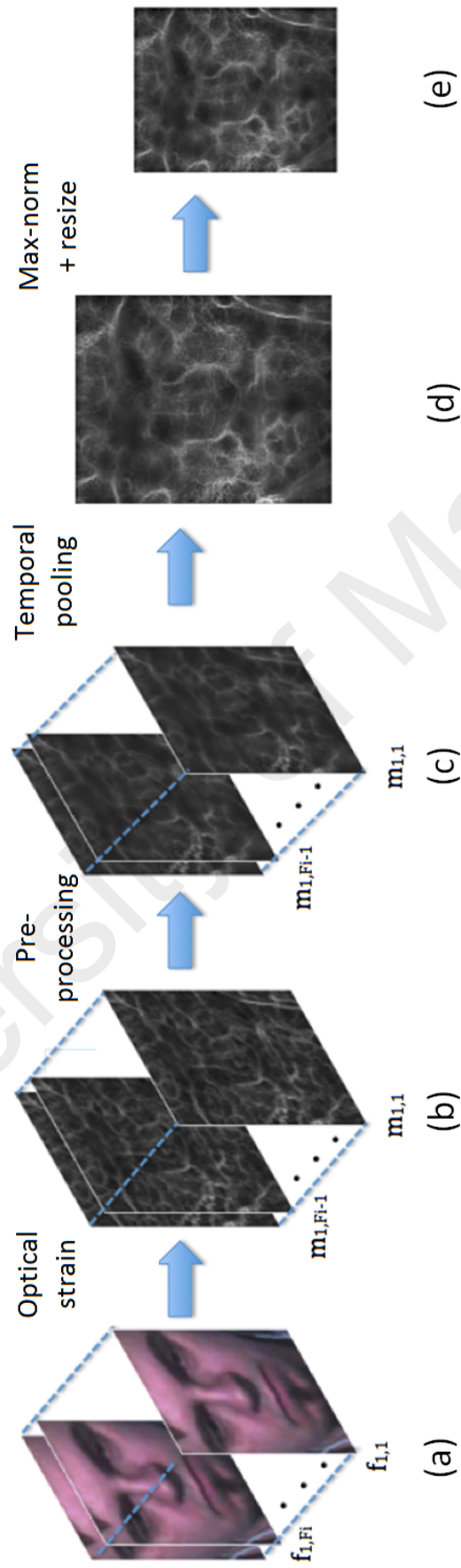
Figure 4.2 illustrates the effect of  $\tau_l$  and  $\tau_u$  on the micro-expression recognition rate. It is observed that  $\tau_l = \tau_u = 0.05$  yields the best results. Therefore, the clipping tolerance is set to 5% of the magnitude range of each processed frame.

With each frame properly aligned, the optical strain maps can then be segmented vertically into three regions of equal size (i.e. forehead–lower eyelid, lower eyelid–nostril and nostril–mouth) to obtain their individual local threshold values. The purpose of performing this segmentation step is to minimize the effects of dominant motions that arise from a particular region as the range of strain magnitudes differ across the three regions. Figure 4.3 shows how an optical strain map is divided into the three vertical segments.





**Figure 4.3:** Example of vertical segmentation of the optical strain frame into three regions



**Figure 4.4:** Extracting OSF from a sample video sequence: (a) Original images,  $f_{1,j}$ ; (b) Optical strain maps  $m_{1,j}$ ; (c) Images after pre-processing; (d) Temporal pooled strain image; (e) Normalized and resized strain image

#### 4.3.1.2 Extracting Optical Strain Features

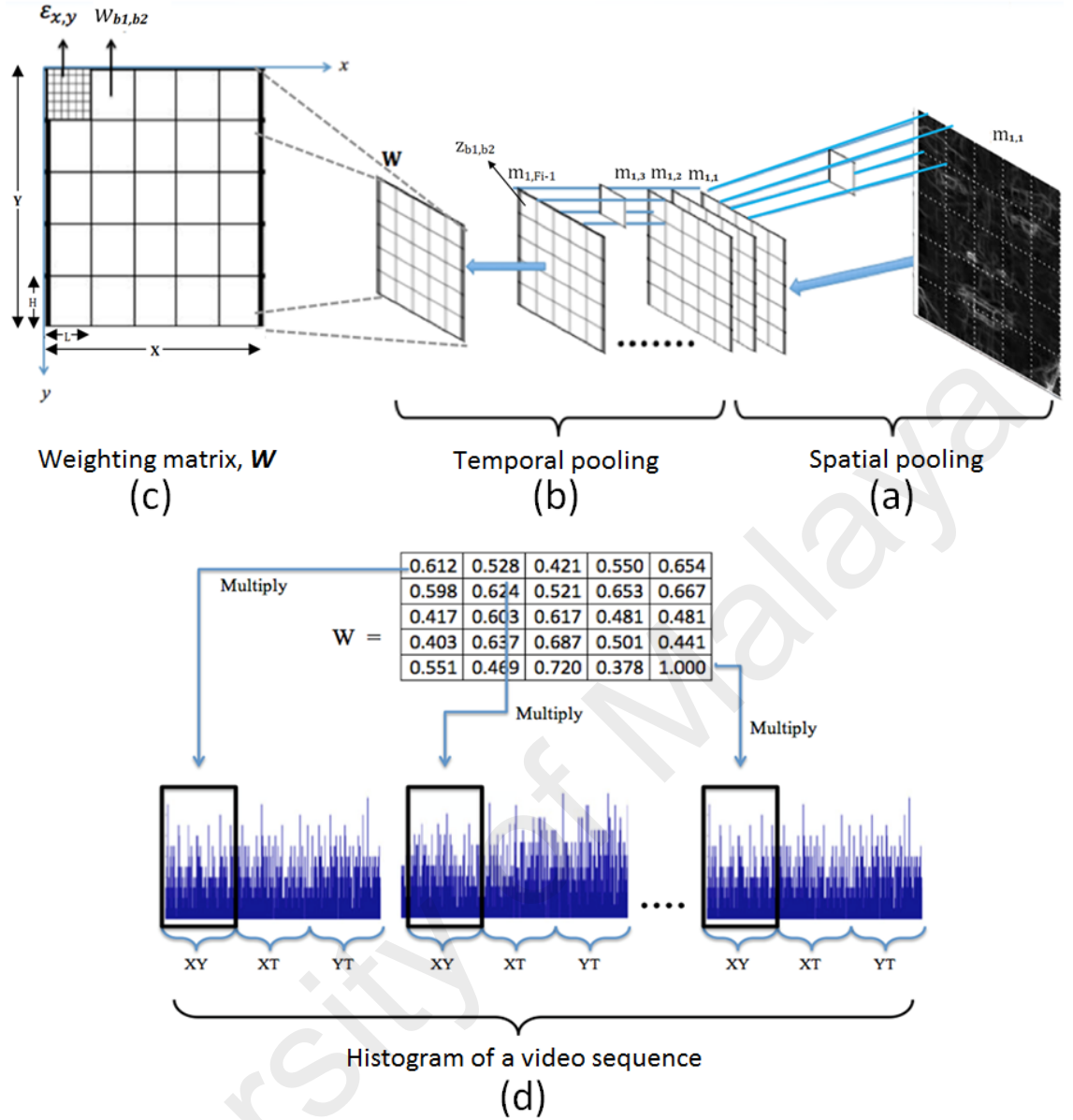
In order to describe the optical strain patterns in a compact and consistent representation, the optical strain maps  $m_{i,j}$  are pooled across time (i.e. temporal pooling). Temporal mean pooling is performed to obtain a composite strain map:

$$\hat{m}_i = \frac{1}{F_i - 1} \sum_{j=1}^{F_i-1} m_{i,j}, \quad (4.5)$$

where all optical strain magnitudes  $\epsilon_{x,y}$  for each strain map  $m_{i,j}$  are averaged across the temporal dimension. The intuition behind this pooling step is to help in accentuating the minute motions in micro-expressions by aggregation of these facial strain patterns. Mean pooling also ensures that the optical strain magnitudes are normalized based on their respective sequence lengths. Then, the composite strain map is max-normalized to increase the significance of its values. In the final step, the composite strain map is resized to  $50 \times 50$  pixels and vectorized its rows to form a 2500-dimension feature vector. Figure 4.4 shows a graphical illustration of the entire process of extracting optical strain features.

#### 4.3.2 Optical Strain Weighted Features

While the OSF describes pixel-level motion features, the LBP-TOP is capable of encoding texture dynamics in larger facial patches. In block-based LBP-TOP (Zhao & Pietikainen, 2007), the feature histograms obtained from all blocks are given equal treatment. Since subtle expressions typically occur in highly localized regions of the face (and this differs for different expression classes), the feature histogram representing these regions should be amplified. As such, larger motions will generate larger optical strain magnitudes and vice versa. A set of weights are then computed to scale the features in each block proportionally to their respective motion strengths. The proposed procedures to obtain Optical Strain Weighted Features (OSW) are: (a) extracting block-based LBP-TOP features; (b)



**Figure 4.5:** OSW histogram formation: (a) Each  $j$ -th frame in  $m_{1,j}$  is divided into  $5 \times 5$  blocks before the values of  $\epsilon_{x,y}$  within each block region are spatially pooled; (b) The block-wise strain magnitudes  $z_{b1,b2}$  from all frames ( $j \in 1 \dots F_{i-1}$ ) are temporally mean pooled; (c) The weighting matrix  $W$  of size  $N \times N$  is formed; (d) Coefficients of  $W$  are multiplied by their respective  $XY$ -plane histogram bins

pre-processing to remove image noises; (c) spatio-temporal pooling on the optical strain maps and determining the weights; (d) multiplying the optical strain weightage on LBP-TOP features. The entire process of obtaining the OSW histogram is graphically shown in Figure 4.5. Details of the entire process are elaborated in the following sub-chapters.

#### 4.3.2.1 Extracting Block-based LBP-TOP Features

Features are extracted by block-based LBP-TOP from each video clip  $s_i$ , whereby the entire video volume is partitioned into  $N \times N$  non-overlapping block volumes. For each block volume, LBP features are computed from three orthogonal planes (concatenated to form LBP-TOP) to obtain dynamic texture features that are local to each particular block region. Finally, the feature histograms from all  $N \times N$  block volumes are concatenated to form the final feature histogram.

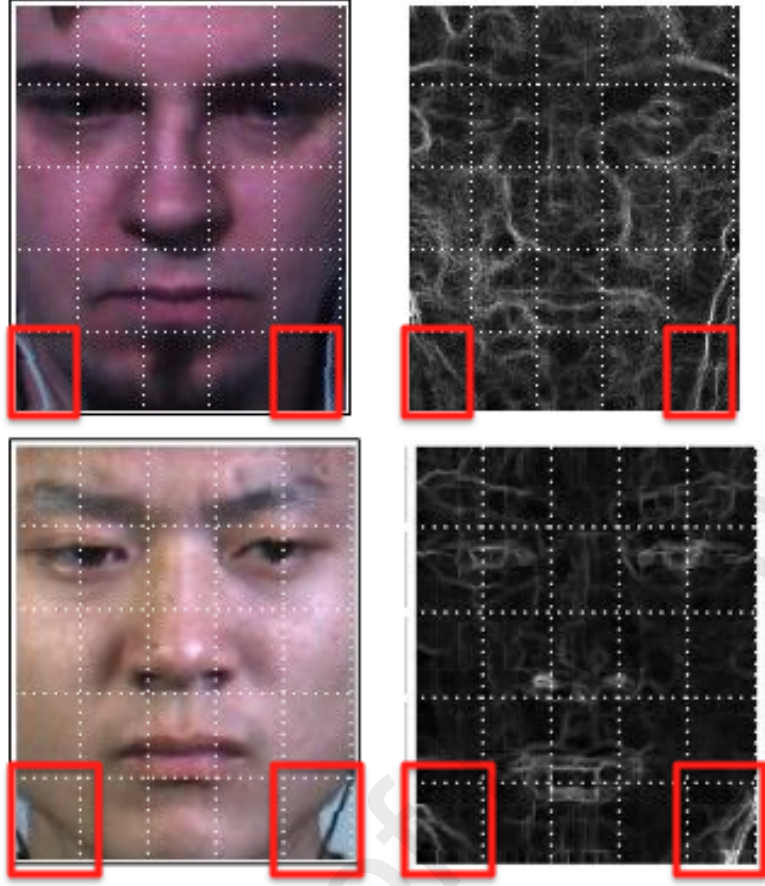
#### 4.3.2.2 Pre-processing

Upon partitioning into blocks, two blocks that are located at the left and right bottom corner of the frames are eliminated due to noticeable amount of movements or noises caused by the background lighting condition, and also the presence of wires from the headset worn by the subjects (see Figure 4.6 for a frame sample from both SMIC-HS and CASME II datasets). For simplicity, these two blocks are referred to as *noise blocks*. Therefore, these noise blocks are omitted and only the remaining  $N^2 - 2$  blocks are utilized for building the feature histogram.

Gaussian noise is the most common noise acquired unintentionally during the elicitation of the micro-expression videos. It may be caused by non-uniform illumination or flickering lights captured by the high speed camera. Since the motions characterized by the subtle facial expressions are fine, it is likely that the Gaussian noises might be incorrectly identified as fine facial motions. Thus, all the images are filtered by a Gaussian filter to reduce noise. The filter size applied is  $5 \times 5$  pixels with standard deviation of  $\sigma = 0.5$  in order to reduce the existing background noises prior to processing.

#### 4.3.2.3 Determining Weights by Spatio-temporal Pooling

To obtain the weights for each block, spatio-temporal pooling is performed on all optical strain maps  $m_{i,j}$  in the video sequence. Spatio-temporal pooling is considered in a sepa-



**Figure 4.6:** Top row: a sample image from SMIC-HS (left) and the corresponding optical strain map (right). Bottom row: a sample image from CASME II (left) and the corresponding optical strain map (right). Noise block at the bottom left and right corners are marked

erable fashion, where spatial mean pooling is performed first, followed by temporal mean pooling.

Firstly, spatial mean pooling averages all the strain magnitudes  $\epsilon_{x,y}$  within each block, resulting in a block-wise strain magnitude:

$$z_{b_1,b_2} = \frac{1}{HL} \sum_{y=(b_2-1)H+1}^{b_2H} \sum_{x=(b_1-1)L+1}^{b_1L} \epsilon_{x,y}, \quad (4.6)$$

where  $L = \frac{X}{N}$ ,  $H = \frac{Y}{N}$ , the block indices  $(b_1, b_2) \in 1, 2, \dots, N$ , and  $(X, Y)$  are the dimensions (width and height) of the frame. This process summarizes the encoded features locally in each block area of the face. Figure 4.5(a) illustrates the spatial pooling process on a strain map,  $m$ .

Secondly, temporal mean pooling is applied on the spatially-pooled frames, where the values of  $z_{b_1, b_2}$  are averaged along the temporal axis across all video frames. The temporal pooling process is illustrated in Figure 4.5(b). Therefore, for each video, a unique set of  $N \times N$  weights is derived,  $\mathbf{W}_i = \{w_{b_1, b_2}\}_{b_1, b_2=1}^N$  (see Figure 4.5(c)), where each weight coefficient  $w$  is defined as:

$$\begin{aligned} w_{b_1, b_2} &= \frac{1}{F_i - 1} \sum_{t=1}^{F_i-1} z_{b_1, b_2} \\ &= \frac{1}{(F_i - 1)HL} \sum_{t=1}^{F_i-1} \sum_{y=(b_2-1)H+1}^{b_2H} \sum_{x=(b_1-1)L+1}^{b_1L} \varepsilon_{x,y}. \end{aligned} \quad (4.7)$$

#### 4.3.2.4 Weighted XY-Plane Histogram

After obtaining the feature histograms extracted by LBP-TOP and the optical strain weights, the coefficients of the weight matrix  $\mathbf{W}$  are multiplied with the XY-plane feature histograms of their corresponding matching blocks. This weighting procedure is performed only on features from the XY-plane so that the motion strengths are well accentuated in each local area of the face as shown in Figure 4.5(d).

Specifically, the optical strain weighted histograms can be defined as:

$$G_{b_1, b_2, d, c} = \begin{cases} w_{b_1, b_2} \bar{M}_{b_1, b_2, d, c}, & \text{if } d = 0; \\ \bar{M}_{b_1, b_2, d, c}, & \text{otherwise,} \end{cases} \quad (4.8)$$

where  $\bar{M}$  is the normalized feature histogram for the block  $(b_1, b_2)$  from Equation (2.7).

#### 4.3.3 Concatenating OSF and OSW Features

In the final step, the two extracted features, namely OSF and OSW features, are concatenated into a single composite feature histogram, named as OSF + OSW. The concatenation process enriches the variety of features used, providing further robustness towards

the detection and recognition of facial micro-expressions. The dimension of the feature histogram in LBP-TOP with  $5 \times 5$  block partitions are  $5 \times 5 \times 3 \times 15$  (OSW) +  $50 \times 50$  (OSF) = 3,625 per video.

## 4.4 Experiments

### 4.4.1 Datasets

The experiments are carried out on two high speed micro-expression databases (i.e., SMIC-HS and CASME II (Yan, Li, et al., 2014; Li et al., 2013)), which are exactly the same databases considered in Chapter 3. Note that, all the image data from these prior-databases are captured under constrained laboratory condition and have undergone face registration and alignment. The methods proposed are evaluated on two separate experiments: (a) detection of micro-expressions (SMIC-HS only), and (b) recognition of micro-expressions (CASME II and SMIC-HS). The detection task determines whether any micro-expression is present. Meanwhile, the recognition task identifies the emotional state that presents in the video clip. Since CASME II does not provide non-micro-expression videos, the detection task is not conducted for this database.

### 4.4.2 Setup

Note that both CASME II and SMIC-HS databases provide the cropped face video sequence, where only the face region is retained while the unnecessary background have been removed. The cropped image frames are directly used in our experiments. These frames have an average spatial resolution of  $340 \times 280$  for CASME II and  $170 \times 140$  pixels for SMIC-HS. The parameter setting used in the experiments are established here. Specifically, the parameters for the feature extractor and classifier are mostly the same values as in the original work, i.e., CASME II (Yan, Li, et al., 2014) and SMIC-HS (Li et al., 2013).

In the detection and recognition tasks performed on the CASME II and SMIC-HS



databases, SVM with linear kernel ( $c = 10000$ ) is utilized as classifier. For the block sizes of LBP-TOP, they are selected based on the original works (Yan, Li, et al., 2014; Li et al., 2013), where the block partitions are  $5 \times 5$  and  $8 \times 8$ . In addition, the number of neighbouring points and the radii along the three orthogonal planes are set to  $LBP-TOP_{4,4,4,1,1,4}$ . The reason of selecting these parameter configurations is explained in Chapter 4.5.2.

However, there are some slight differences in the SVM protocol settings for the two databases. Specifically, the recognition task on CASME II is a five-class problem (i.e., disgust, happiness, tense, surprise and repression). It is evaluated using SVM classifier with LOVOCV setting, adopted from the original work (Yan, Li, et al., 2014). On the other hand, in SMIC-HS, the recognition task is a three-class classification (i.e., positive, negative, surprise classes), while detection of micro-expressions is a binary decision (i.e., yes / no). Evaluations on SMIC-HS are conducted using SVM classifier with LOSOCV setting, following the original work (Li et al., 2013).

There are two ways to measure the classification performance in the LOSOCV setting, namely *macro*- and *micro*-averaging (Tsoumakas et al., 2010). The macro-averaged result gives the accuracy computed by averaging across all individual subject-wise accuracy results. On the other hand, micro-averaged result refers to the overall accuracy result across all evaluated samples. Further performance metrics are also presented, such as F-measure, precision and recall when the LOSOCV setting is employed. These three metrics provide a more meaningful perspective than accuracy rates when the datasets used are naturally imbalanced since each subject has a different number of video frames. Refer to Chapter 3.5.2 for the derivation of these metrics.

The three proposed methods: (a) OSF, (b) OSW, and (c) OSF + OSW, are then evaluated and compared to their respective baseline methods (Yan, Li, et al., 2014; Li et al., 2013) on both detection and recognition experiments.

**Table 4.1:** Micro-expression detection and recognition results on SMIC-HS and CASME II database with LBP-TOP of  $5 \times 5$  block partitioning

Methods	Detection - SMIC-HS		Recognition - SMIC-HS		Recognition CASME II
	Micro-avg	Macro-avg	Micro-avg	Macro-avg	
Baseline	.61	.66	.41	.43	.62
OSF	.66	.66	.42	.46	.51
OSW	.64	.68	.47	.49	.63
OSF + OSW	<b>.72</b>	<b>.75</b>	<b>.44</b>	<b>.48</b>	<b>.63</b>

**Table 4.2:** Micro-expression detection and recognition results on SMIC-HS and CASME II database with LBP-TOP of  $8 \times 8$  block partitioning

Methods	Detection - SMIC-HS		Recognition - SMIC-HS		Recognition CASME II
	Micro-avg	Macro-avg	Micro-avg	Macro-avg	
Baseline	.57	.59	.46	.48	.61
OSF	.66	.66	.41	.46	.51
OSW	.63	.63	.49	.51	.62
OSF + OSW	<b>.73</b>	<b>.74</b>	<b>.52</b>	<b>.58</b>	<b>.62</b>

## 4.5 Results and Discussions

### 4.5.1 Detection and Recognition Results

From the results shown in Table 4.1 and Table 4.2, it is observed that the OSF method is capable of producing reasonably positive results compared to the baselines in some cases. However, better and more consistent results are obtained using OSW and OSF + OSW methods for both macro- and micro-averaging measures.

For the detection task on SMIC-HS database (using  $5 \times 5$  blocks in LBP-TOP), OSF + OSW outperforms the baseline by 11% and 9% in micro- and macro-averaged results respectively, as shown in Table 4.1. In addition, more significant improvement of 16% (for micro-averaged) is achieved, when the block partition of  $8 \times 8$  is used, as tabulated in Table 4.2. Furthermore, Table 4.3 lists the confusion matrices of the detection results on SMIC-HS database. The proposed method OSF + OSW is able to better distinguish the micro- and non-micro-expressions. It can be seen that for non micro-expression, there is a significant increase in recognition rate of approximately 17%.

In the recognition experiment on the SMIC-HS database, it is able to achieve up

**Table 4.3:** Confusion matrices of baseline and OSF + OSW methods for detection task on SMIC-HS database with LBP-TOP of  $5 \times 5$  block partitioning

(a) Baseline		
	micro-expression	non-micro-expression
micro-expression	<b>.65</b>	.35
non-micro-expression	.43	<b>.57</b>

(b) OSF + OSW		
	micro-expression	non-micro-expression
micro-expression	<b>.70</b>	.30
non-micro-expression	.26	<b>.74</b>

to 5% improvement (for macro-averaging) over the baseline results on the concatenated OSF + OSW method using  $5 \times 5$  in LBP-TOP (see Table 4.1). This method also registers a performance improvement of 10% (also for macro-averaging) over the baseline results when  $8 \times 8$  block partition is used (see Table 4.2). These results point towards a significant improvement in feature representation when optical strain information is well-utilized. It is worth noting that although the OSF method did not perform as well, its contribution towards the concatenated OSF + OSW method should not be disregarded. The detailed confusion matrices for the recognition performance on SMIC-HS database utilizing the baseline and OSF + OSW methods are shown in Table 4.4. It can be seen that ‘Negative’ and ‘Surprise’ expressions can be recognized with higher accuracy when using OSF + OSW method, while the accuracy of the ‘Positive’ expression remains unchanged at 49%.

On the other hand, for the recognition experiment on the CASME II dataset, it is observed from Table 4.1 and Table 4.2 that there is a substantial improvement in OSF + OSW method for both the  $5 \times 5$  and  $8 \times 8$  block partitions in LBP-TOP. Table 4.5 shows the confusion matrices for the recognition results on CASME II database. It can be seen that ‘Disgust’, ‘Tense’ and ‘Surprise’ are recognized with higher accuracy using the OSF + OSW method, but ‘Happiness’ and ‘Repression’ have lower recognition rate compared to the baseline. However, the average recognition accuracy for the OSF + OSW method

**Table 4.4:** Confusion matrices of baseline and OSF + OSW methods for recognition task on SMIC-HS database with LBP-TOP of  $8 \times 8$  block partitioning

(a) Baseline				(b) OSF + OSW			
	Negative	Positive	Surprise		Negative	Positive	Surprise
Negative	<b>.43</b>	.36	.21	Negative	<b>.46</b>	.36	.19
Positive	.37	<b>.49</b>	.14	Positive	.39	<b>.49</b>	.12
Surprise	.33	.21	<b>.47</b>	Surprise	.23	.09	<b>.67</b>

**Table 4.5:** Confusion matrices of baseline and OSF + OSW methods for recognition task on CASME II database with LBP-TOP of  $5 \times 5$  block partitioning

(a) Baseline						
	Disgust	Happiness	Tense	Surprise	Repression	
Disgust	<b>.55</b>	.05	.30	.08	.02	
Happiness	.06	<b>.55</b>	.15	0	.24	
Tense	.23	.05	<b>.73</b>	0	0	
Surprise	.24	.08	.20	<b>.44</b>	.04	
Repression	.04	.22	.11	0	<b>.63</b>	

(b) OSF + OSW						
	Disgust	Happiness	Tense	Surprise	Repression	
Disgust	<b>.68</b>	.07	.18	.07	0	
Happiness	.12	<b>.39</b>	.30	.03	.15	
Tense	.17	.07	<b>.74</b>	.02	.01	
Surprise	.20	.08	.20	<b>.48</b>	.04	
Repression	.11	.22	.11	0	<b>.56</b>	

is better than that of the baseline.

Other performance metrics (including F-measure, recall and precision) are reported in Table 4.6 and Table 4.7. The two tables further substantiate the superiority of the proposed OSF + OSW method over the baseline method in both the detection and recognition tasks. For CASME II recognition task, the performance of the concatenated OSF + OSW method appears to be as good as that achieved by the baseline.

#### 4.5.2 Discussions

In a nutshell, optical strain characterizes the relative amount of displacement by a moving object within a time interval. Its ability to capture any small muscular movement on face can be advantageous to subtle expression research. By simple product of the LBP-TOP

**Table 4.6:** F-measure, recall and precision scores for detection and recognition performance on SMIC-HS and CASME II database with LBP-TOP of  $5 \times 5$  block partitioning

(a) Detection - SMIC-HS

Methods	Micro-averaging			Macro-averaging		
	F-measure	Recall	Precision	F-measure	Recall	Precision
Baseline	.61	.61	.61	.66	.66	.67
OSF	.66	.66	.66	.67	.66	.69
OSW	.64	.64	.64	.69	.68	.71
OSF + OSW	<b>.72</b>	<b>.72</b>	<b>.72</b>	<b>.77</b>	<b>.75</b>	<b>.79</b>

(b) Recognition - SMIC-HS

Methods	Micro-averaging			Macro-averaging		
	F-measure	Recall	Precision	F-measure	Recall	Precision
Baseline	.41	.41	.40	.37	.38	.39
OSF	.42	.42	.41	.38	.38	.42
OSW	.47	.47	.47	.45	.44	.49
OSF + OSW	<b>.45</b>	<b>.46</b>	<b>.44</b>	<b>.37</b>	<b>.37</b>	<b>.41</b>

(c) Recognition - CASME II

Methods	F-measure	Recall	Precision
Baseline	.59	.58	.61
OSF	.44	.41	.47
OSW	.62	.61	.64
OSF + OSW	<b>.59</b>	<b>.57</b>	<b>.61</b>

histogram bins with the weights, the resulting feature histograms are intuitively scaled to accommodate the importance of block regions. The OSF + OSW approach generates consistently promising results throughout the experiments tested on the SMIC-HS database. The reason why the proposed method did not performed as good on the CASME II dataset as the experiments conducted on SMIC-HS, is probably due to the high frame rate of its acquired video clips. Repetitive frames with very little changes in movements might result in redundancy of input data. Hence, the extracted strain information may be insignificant (hence negligible) to offer much discrimination between features of different classes. This is most obvious in the OSF results for CASME II, where there is in fact a significant deterioration of performance. SMIC-HS videos, on the other hand (at only half the frame rate of CASME II videos), is able to harness the full capability of optical strain information where the OSF is seen to complement OSW very well, producing even

**Table 4.7:** F-measure, recall and precision scores for detection and recognition performance on SMIC-HS and CASME II database with LBP-TOP of  $8 \times 8$  block partitioning

(a) Detection - SMIC-HS

Methods	Micro-averaging			Macro-averaging		
	F-measure	Recall	Precision	F-measure	Recall	Precision
Baseline	.57	.57	.57	.60	.59	.62
OSF	.66	.66	.66	.67	.66	.69
OSW	.63	.63	.63	.65	.63	.67
OSF + OSW	<b>.73</b>	<b>.73</b>	<b>.73</b>	<b>.77</b>	<b>.74</b>	<b>.80</b>

(b) Recognition - SMIC-HS

Methods	Micro-averaging			Macro-averaging		
	F-measure	Recall	Precision	F-measure	Recall	Precision
Baseline	.46	.46	.46	.41	.423	.42
OSF	.42	.42	.41	.38	.38	.42
OSW	.50	.51	.50	.42	.43	.43
OSF + OSW	<b>.53</b>	<b>.54</b>	<b>.53</b>	<b>.46</b>	<b>.46</b>	<b>.48</b>

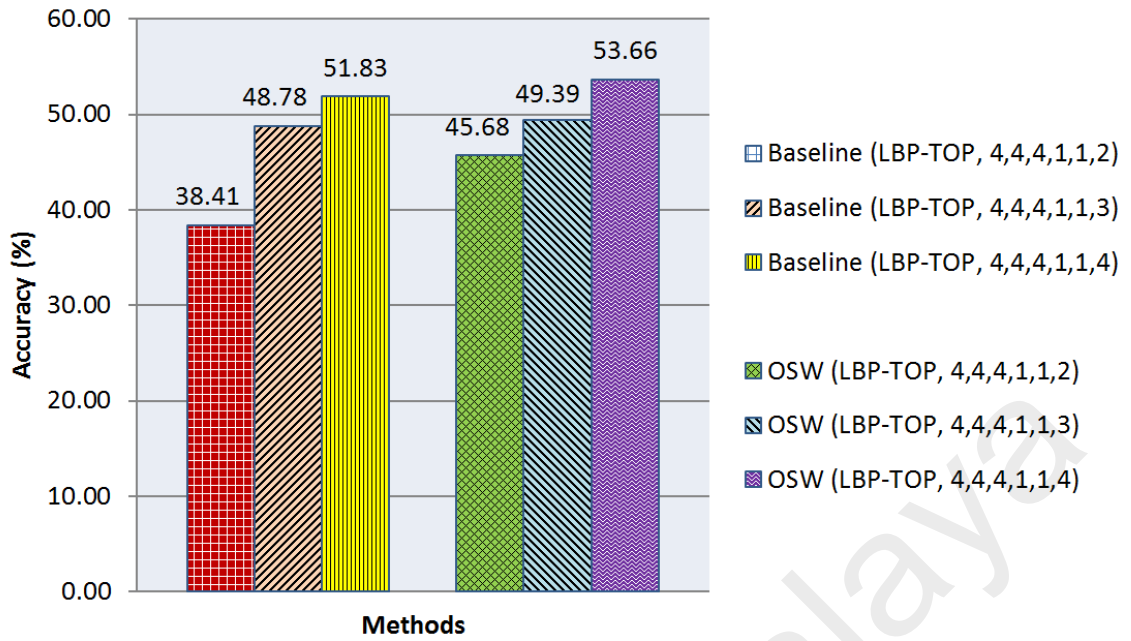
(c) Recognition - CASME II

Methods	F-measure	Recall	Precision
Baseline	.58	.56	.60
OSF	.44	.41	.47
OSW	.59	.57	.61
OSF + OSW	<b>.57</b>	<b>.56</b>	<b>.59</b>

better results when combined together.

Since there are background noises in the video frames from both databases, spatial pooling helps to improve the robustness against these noises. Furthermore, high strain magnitudes detected in the frame that exceeded the upper threshold (Equation (4.4)) are treated as noises and not micro-expression movements. On the other hand, low strain magnitudes below the lower threshold (Equation (4.3)) will also be ignored since they do not contribute sufficient details towards the micro-expressions.

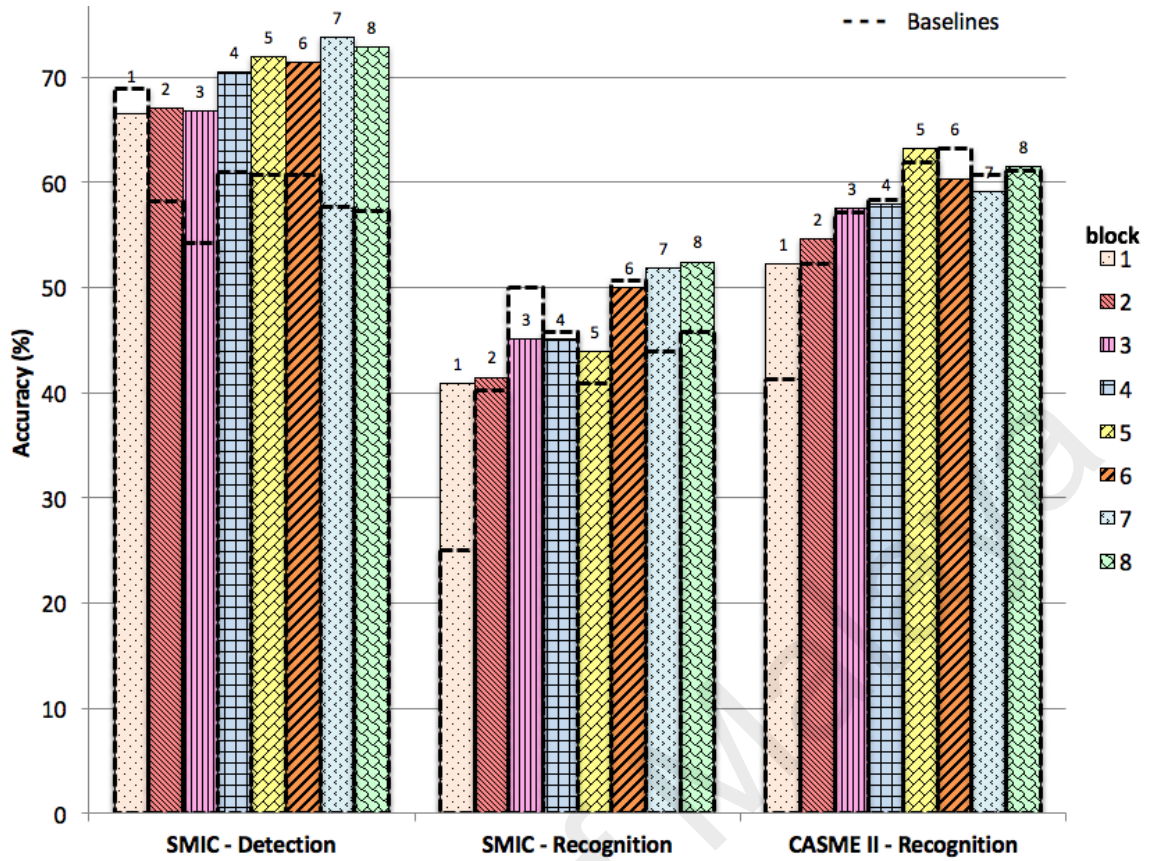
Another notable observation worth mentioning lies with the radii parameters of the LBP-TOP feature extractor (which is used by OSW method). As shown in Figure 4.7, by varying the value of  $R_T$  (temporal radius), the recognition accuracy is the highest for both the OSW and baseline (LBP-TOP) methods when  $R_T = 4$ . Therefore, all the OSW experiments on the SMIC-HS database are conducted using  $LBP-TOP_{4,4,4,1,1,4}$  to



**Figure 4.7:** Micro-averaged accuracy results of the baseline (LBP-TOP) and OSW methods using different LBP-TOP radii parameters on SMIC-HS database based on LOSOCV

maximize the performance on accuracy.

The block partition settings that are used in the original papers of CASME II (Yan, Li, et al., 2014) and SMIC-HS (Li et al., 2013) are applied. Specifically, the detection task in SMIC-HS uses  $5 \times 5$  blocks, while the recognition tasks in SMIC-HS and CASME II use  $8 \times 8$  and  $5 \times 5$  blocks, respectively. Figure 4.8 demonstrates in detail how the OSF + OSW fare with different block size configuration, with the baselines indicated by dashed lines. It can be seen that the larger blocks (i.e. smaller number of partitions,  $N = 1, 2, 3$ ) do not produce better results compared to smaller blocks (i.e. larger number of partitions,  $N = 6, 7, 8$ ) in all scenarios. This is because the local facial appearance and motion that carry important details at specific facial locations are not well described in large block areas. Hence, this analysis justifies our choice of using the block settings suggested in the original works, where the best results using the block-based LBP-TOP feature can be achieved. On the other hand, it is also clear in Figure 4.8 that the proposed OSF + OSW method outperforms the baseline LBP-TOP (dashed lines with transparent fill) in a majority of the experiments.



**Figure 4.8:** Recognition accuracy results of the baseline (LBP-TOP) and OSF + OSW methods using different block partitions in LBP-TOP. The baseline results are denoted by the dashed lines (with transparent fill)

#### 4.5.3 Comparison with Other Spatio-temporal Features

The accuracy for the detection and recognition tasks are reported in Table 4.8. The proposed OSF + OSW (method #13) method is compared against other spatio-temporal based features, namely: (a) the optical flow based features OFF + OFW (method #12), which is constructed in the similar manner (optical flow magnitudes used instead of optical strain magnitudes); (b) STIP (method #2) or Histogram of Oriented Gradients and Histogram of Optical Flow, extracted from spatio-temporal interest points (H. Wang et al., 2009), and; (c) HOG3D (method #3) or 3D Oriented Gradients (Klaser et al., 2008). The last two descriptors (i.e., (b) and (c)) are popular spatio-temporal features used in various human action recognition (Kovashka & Grauman, 2010) and facial expression recognition analysis (Hayat et al., 2012). For both of these methods, the interest points are densely sampled with their default parameters specified by the authors, and bag-of-words (BOW)



(H. Wang et al., 2009) representation is used to learn the visual vocabulary and build the feature vectors. The number of clusters or “bags” used in the vocabulary learning is determined empirically and the best result is reported. For all these methods, SVM classifier with linear kernel is applied for fair comparison, except for the methods TICS (method #8) (S. Wang et al., 2015) and MDMO (method #9) (Y. J. Liu et al., 2016), where they classified the micro-expression in CASME II into four categories (i.e., negative, positive, surprise and others), instead of five (i.e., disgust, happiness, tense, surprise and repression). Besides, MDMO utilized polynomial kernel in SVM with heuristically determined parameter settings.

In Table 4.8, it is observed that STIP and HOG3D features yielded poor results because they are not designed to capture fine appearance and motion changes. On the other hand, the performance of OFF + OFW features are more comparable to the baseline performance of the SMIC-HS, but it is inferior to the CASME II baseline by a significant amount. Overall, the proposed OSF + OSW features yielded promising detection and recognition results compared to the other spatio-temporal features evaluated. It can be concluded that the proposed method is capable of describing the spatio-temporal information in micro-expressions in a more effective manner.

**Table 4.8:** Comparison of micro-expression detection and recognition accuracy results on the SMIC-HS and CASME II databases for different feature extraction methods

#	Methods	Detection - SMIC-HS*		Recognition - SMIC-HS*		Recognition CASME II*
		Micro-averaging	Macro-averaging	Micro-averaging	Macro-averaging	
1	Baselines <sup>†</sup>	.61	.66	.46	.48	.62
2	STIP (H. Wang et al., 2009)	.55	.59	.42	.41	.47
3	HOG3D (Klaser et al., 2008)	.61	.63	.41	.37	.51
4	RW (Oh et al., 2015)	N/A	N/A	.34	N/A	N/A
5	STM (Le Ngo et al., 2014)	N/A	N/A	.44	N/A	N/A
6	LBP-SIP (Y. Wang et al., 2014)	N/A	N/A	.55	N/A	.67
7	STLBP-IP (X. Huang et al., 2015)	N/A	N/A	.60	N/A	N/A
8	TICS <sup>♦</sup> (S. Wang et al., 2015)	N/A	N/A	N/A	N/A	.62
9	MDMO <sup>♦</sup> (Y. J. Liu et al., 2016)	N/A	N/A	N/A	N/A	.70
10	OSF	.66	.66	.41	.46	.51
11	OSW	.63	.63	.49	.51	.62
12	OFF + OFW	.62	.61	.40	.42	.56
13	OSF + OSW	<b>.73</b>	<b>.74</b>	<b>.52</b>	<b>.58</b>	<b>.63</b>

<sup>†</sup> Baseline results from (Li et al., 2013; Yan, Li, et al., 2014)

<sup>♦</sup> Used 4 classes for CASME II instead of 5

\* LOSOCV protocol in SVM

• LOVOCV protocol in SVM

## 4.6 Summary

A novel feature extraction approach is proposed for the detection and recognition of facial micro-expressions in video clips, which is mainly built on the optical strain technique. The proposed method describes the fine subtle movements on the face using optical strain in two different ways. The first, OSF, is a direct utilization of optical strain information as a feature histogram. Secondly, OSW is the utilization of strain information as weighted coefficients to LBP-TOP features. Lastly, OSF + OSW is the concatenation of the two feature histograms to form the resultant feature histogram. The viability of optical strain information is demonstrated in the experiments by considering two recent state-of-the-art micro-expression databases, namely SMIC-HS and CASME II. More importantly, the proposed OSF + OSW feature descriptor is capable of achieving promising results in both the detection and recognition tasks.

The best detection performance for SMIC-HS is 75% using  $5 \times 5$  block partition in LBP-TOP, which translates to a significant improvement of more than 9% compared to the baseline results. The improvement is more remarkable when  $8 \times 8$  block partitions are used, where OSF + OSW method is able to achieve a maximum improvement of approximately 15%. In addition, OSF + OSW method is able to attain an improvement of +5% and +10% on micro-expression recognition performance for the SMIC-HS dataset using  $5 \times 5$  and  $8 \times 8$  block partitions, respectively. The aforementioned results tested on the SMIC-HS dataset are the macro-averaged performance by using SVM with linear kernel. On the other hand, results of OSF + OSW method on the CASME II are slightly better when compared to that of the baselines, i.e., +1% for both the  $5 \times 5$  and  $8 \times 8$  block partitions.

There are many avenues for further research. The kernel function used in SVM is quite sensitive towards the given data, and how this can be better chosen can be further

studied. In addition, better noise filtering techniques and masking of different face regions can be applied to alleviate the instability of illumination and intensity changes on the face areas or background. This can potentially help in reducing the erroneous optical flow and optical strain computation.

University of Malaya

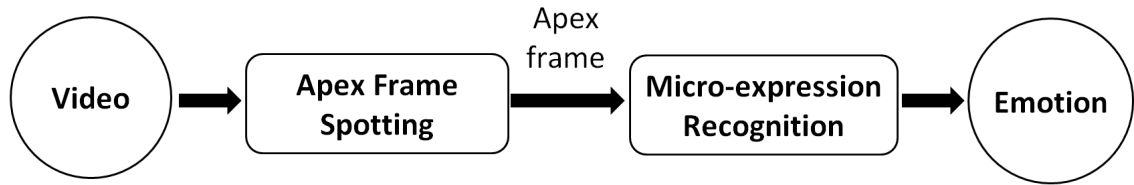
## CHAPTER 5: FEATURE EXTRACTION USING APEX FRAME

### 5.1 Overview

In this chapter, novel automatic approaches for micro-expression recognition that combine both the spotting and recognition mechanisms, are introduced. More precisely, two spotting mechanisms are designed separately to examine short and long video datasets. This is because the attributes of the short and long videos are different, as detailed in Chapter 2.4.2. To recap, the short videos only comprise micro-expression frames (i.e., from onset to offset), whereas the long videos might contain irrelevant micro-expression movements, such as eye blinking action, which can possibly lead to inaccuracy in apex frame spotting.

Nevertheless, the basic flow of the entire proposed micro-expression recognition mechanisms for both the short and long videos are the same, as illustrated in Figure 5.1. Specifically, in the apex frame spotting task, the index of the apex frame is first identified from the entire video sequence. The reason of spotting only the apex frame is that it is the frame that contains the highest intensity of expression changes among all frames. In contrast, the first frame of each video is assumed to be the neutral frame, which has the least expression changes. Next, for the recognition task, a new feature descriptor, Bi-Weighted Oriented Optical Flow (Bi-WOOF) is proposed, to encode facial micro-expression features by utilizing only the apex frame (and the first frame as reference frame) among all frames of an entire video sequence.

Experiment results suggest that the proposed approach is sufficient to produce high recognition accuracy. To date, this is the first attempt at recognizing micro-expressions from videos using only the apex frame. The proposed algorithms are evaluated on four short video micro-expression databases (i.e., CASME II, SMIC-HS, SMIC-NIR and SMIC-



**Figure 5.1:** Flowchart of the apex frame spotting and emotion recognition system

VIS) and four long video databases (i.e., CASME II-RAW, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR). The details of these databases can be found in Chapter 2.4.

## 5.2 Introduction

This chapter studies two primary elements in micro-expression system, notably, the spotting and recognition tasks. Specifically, spotting task is to indicate the interval of micro-expression occurrence or the frame indices of some important instants (such as onset, apex and offset), while recognition task is to classify the expression type (such as disgust, happiness, repression, surprise and tense) given a micro-expression video sequence. In the literature, majority of the articles focused on the recognition analysis, that mainly validate the new approaches on the entire video sequence (that comprises micro-expression frames only).

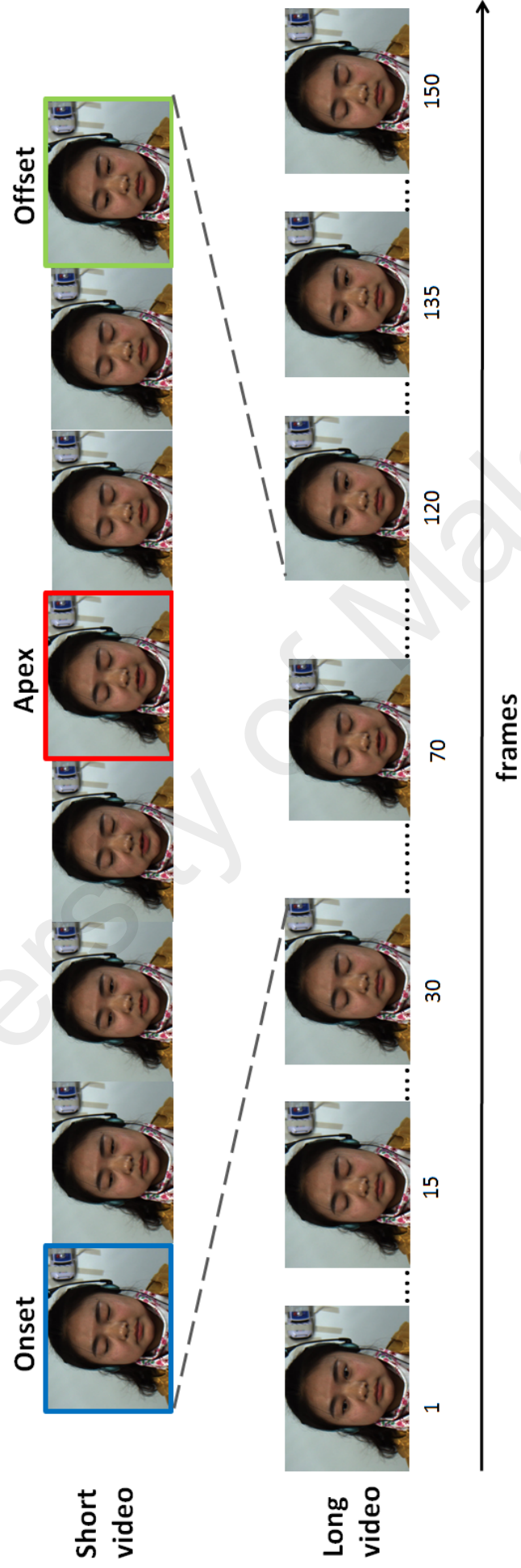
A short recall from Chapter 3.3.1, micro-expression is a dynamic facial action which evolves in the following sequence of states: neutral-onset-apex-offset-neutral. Starting from a neutral state, an onset frame indicates the beginning of a micro-expression where the facial muscles begin to undergo contraction, while offset frame is the end of the expression where the intensity of the muscles is reduced to zero. Apex frame is the instant when the micro-expression reaches its climax (the most intense movement). The apex is not necessarily located at the middle between the onset and offset frames, but it can be situated at any frame between the onset-offset range.

Here, the video sub-sequence that is composed of only the frames from onset to offset is defined as *short video*. On the other hand, *long video* refers to the raw video

sequence which may include the frames with micro-expressions as well as irrelevant motion that are present before the onset and after the offset. Figure 5.2 illustrates the short and long video sequences with onset-apex-offset frame annotations. Notice how a micro-expression sequence (frames 30-120) in a long video can be easily shrouded by frames outside the onset-offset range that contain eye blinks and head rotations (such as in frames 15 and 150).

### 5.3 Motivation

In current literature, most works categorized micro-expressions using the pre-cropped short videos. For these cases, the locations of the onset and offset frames are required. These annotations can be obtained from the ground-truth, which are manually marked and verified by trained psychologists or “coders”. However, the precision of ground-truth labeling is highly dependent on the judgment of the psychologists, who decide the onset and offset locations using frame-by-frame observation (Yan, Li, et al., 2014; Li et al., 2013). As such, the reliability and consistency of the marking are directly affected. As a consequence, imprecise ground-truth information may influence the recognition accuracy of the micro-expression recognition system. This chapter aims to search for the apex frame automatically, to utilize it in the recognition task for the four major benefits: (a) consistency and high reliability of the labeled data; (b) time and cost saving in data labeling, as manual labeling requires intensive human effort and focus; (c) computational simplicity and high efficiency in processing only the apex frame instead of the entire video sequence, and; (d) redundancy elimination in extracting the repetitive frames with very little changes.



**Figure 5.2:** An example of a long and short video with annotated ground-truth labels indicating the onset, apex and offset frames



## **5.4 Literature Review**

The following sub-chapters discuss some recent works published for micro-expression spotting and recognition analysis, as well as the problems that exist in long videos.

### **5.4.1 Apex Spotting in Short Videos**

Yan et al. (Yan, Wang, Chen, et al., 2014) published the first work to automatically spot the instance of the single apex frame in a video. The micro-expression information retrieved from that apex frame is expected to be insightful in both psychological and computer vision research purposes, because it contains the maximum facial muscle movements throughout the video sequence. They employed two feature extractors (i.e., LBP and CLM) and reported the average frame distance between the spotted apex and the ground-truth apex. The frame that has the highest feature difference between the first frame and the subsequent frames is defined to be the apex. However, the CLM feature performed poorly as it is not able to annotate landmark points to a good degree of accuracy. Furthermore, there are two flaws in Yan et al.'s (Yan, Wang, Chen, et al., 2014) work, namely: (a) the average frame distance calculated is not in absolute mean, which lead to incorrect results, and; (b) the method is validated by using only approximately 20% of the video samples in the database (i.e., CASME II), hence the proposed spotting approach is not conclusive and convincing.

### **5.4.2 Micro-expression Spotting in Long and Short Videos**

There are few conventional methods which attempted to spot the frame instant of the micro-facial movements in the database. For instance, the work by Moilanen et al. (Moilanen et al., 2014) searched for the frame indices that contain micro-expressions. Specifically, a Chi-Squared dissimilarity is utilized to calculate the distribution difference between the LBP histogram of the current feature frame and the averaged feature frame. The frames which yield score greater than an empirically determined threshold

are regarded as frames with micro-expression. The short video databases, CASME A and CASME B, and the long video databases, SMIC-E-VIS are considered to evaluate the proposed method. It is able to achieve 52% true positive rate with 30 false positive in CASME A; 66% spotting accuracy with 32 false positive in CASME B, and; spotting accuracy of 71% with 23 false positive in SMIC-VIS-E. During the experiment, the authors tried to mask the eye regions of the face images to avoid the spotting performance from being affected by any eye related events. Yet, masking the eyes did not prevent the eye blinks to be spotted. Thus, they claimed that the eye blinking movement is considered one type of the micro-expressions. Since the eye blink attribute is not detailed in the ground-truth, the frames that contain the eye blinking movements are annotated manually.

A similar approach is carried out by Davison et al. (Davison et al., 2015) to spot the frames that consist of micro-expressions, except that: (a) a denoising method is added before extracting the features, and; (b) the Histogram of Gradient (HOG) is employed instead of LBP. However, the database they tested on is not publicly available. Since the benchmark video sequences used in this paper (Davison et al., 2015) and that in (Moilanen et al., 2014) are different, their performances cannot be compared directly. Similar to the spotting task carried out by Moilanen et al. (Moilanen et al., 2014), the spotted eye blinking frames are also regarded as true positive.

#### **5.4.3 Micro-expression Spotting and Recognition in Long Videos**

To the best of our knowledge, there is only one prior work which combines the spotting and recognition tasks to categorize the type of micro-expression, performed by Li et al. (Li et al., 2015). They extended the work by Moilanen et al. (Moilanen et al., 2014), where after the spotting stage, the spotted micro-expression frames (i.e., those within the onset and offset range) are concatenated to a single sequence for expression recognition. In the recognition task, a motion magnification technique is employed to magnify

the subtle motion so that the micro-expressions are easier to be distinguished. Besides, a new feature extractor, namely, Histograms of Image Gradient Orientation (HIGO), is introduced to suppress the effect of illumination variation. However, the spotting threshold is chosen heuristically (at true positive rate of 74.86%) to obtain the spotted micro-expression sequences which are fed to the recognition component. Although evaluation on the SMIC-E-VIS showed promising results, the reliance on the annotated onset and offset frames, as well as the use of a tunable threshold parameter warrants the need for manual intervention. Besides, the frame rate of SMIC-E-VIS is 25fps, which means that the maximum frame number in a long video sequence is only  $1/5 \text{ s} \times 25 \text{ fps} = 5 \text{ frames}$ .

#### **5.4.4 Eye Blinking Issue in Long Videos**

Evaluation of the micro-expression system on long videos is particularly challenging, primarily because of the presence of unwanted facial movements. These motions correspond to falsely detected micro-expressions, which may appear before the actual onset frame and after the offset frame. One common irrelevant facial movement that is unavoidable during the elicitation of micro-expression database is the eye blinking motion. Shreve et al. (Shreve et al., 2009) suggested to remove the eye regions because eye blinking can adversely affect optical flow estimation, causing false detection of the micro- and macro-expressions. In their work, the boundaries of the eye regions are automatically drawn using a landmark annotator. Unlike the work of Moilanen et al. (Moilanen et al., 2014) and Davison et al. (Davison et al., 2014), the spotted frames that contained the eye blinks are manually marked as true positive.

#### **5.4.5 Feature Extraction and Face Representation**

##### *5.4.5.1 Regions of Interest*

Many research papers demonstrated that extracting the features from certain facial regions improves the facial expression recognition performance, compared to considering the en-

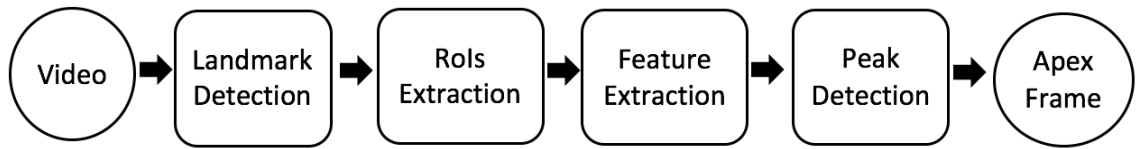
tire face. The main reasons are that, those Regions of Interest (RoIs) contribute more facial changes information towards differentiation of the expressions and can eliminate the parts that do not correspond to the desired facial movements. Happy and Routray (Happy & Routray, 2015) introduced an automated salient facial patches selection method, in which the sub-regions selected depend upon the locations of facial landmarks detected using DRMF. However, there is no commonly agreed standard for specific combination of the facial patches on achieving better accuracy on facial expression analysis. More published works related on utilizing the facial patches are discussed in Chapter 2.2.3.

#### *5.4.5.2 Local Binary Pattern*

To describe the texture of the facial movements, the LBP (Ojala et al., 1996) method, which is a simple operator, is widely used. As a brief review, the intensity value of the center pixel is compared with its neighboring pixels using thresholding technique. The result is encoded in a short binary code to represent the pattern of the neighborhood pixel. Succinctly, LBP has discrimination power, compact representation and low computational complexity. Detailed information has been discussed in Chapter 2.3.1.

#### *5.4.5.3 Optical Strain*

Optical strain can be deployed to measure the intensity of the expression occurred and is therefore applicable for the facial expression detection and recognition tasks. Shreve et al. (Shreve et al., 2014) employed optical strain magnitude to spot micro-expressions and the spotting performances obtained are promising. The capability of optical strain in computing the temporal motion details for each pixel in each frame robustly, justifies its utilization in the micro-expression apex frame spotting application. Chapter 2.3.4 elaborates the advantages of optical strain as a feature descriptor and the related published works.



**Figure 5.3:** Flow diagram of apex frame spotting in short video

## 5.5 Proposed Algorithm

The proposed micro-expression recognition system comprises of two components, namely, apex frame spotting, and micro-expression recognition. The following sub-chapters detail the steps involved for both short and long videos.

### 5.5.1 Apex Frame Spotting in Short Video

The apex frame spotting approach in short videos consists of four steps:

1. Landmark detection - the facial landmark points are first annotated by using a DRMF landmark detector.
2. RoIs extraction - the RoIs that indicate the facial region with important micro-expression details are extracted according to the landmark coordinates.
3. Feature extraction - the LBP feature descriptor is adopted to obtain the features of each frame in the video sequence (i.e., from onset to offset). The feature difference between the first and the rest of the frames are computed using the correlation coefficient formula.
4. Peak detection - a peak detector with *divide-and-conquer* strategy is utilized to search for the apex frame based on the LBP feature difference. Note that all the steps proposed above are fully automated and completely rely on the facial locations marked by the landmark detector.

The process flow diagram of the proposed apex frame spotting approach in short videos is illustrated in Figure 5.3, with detail of each step elaborated as follows.

#### *5.5.1.1 Landmark Detection and RoIs Extraction*

Yan et al. (Yan, Wang, Chen, et al., 2014) reported that the two most expressive facial parts are located in the eyebrow and mouth areas. On the other hand, Ringeval et al. (Ringeval et al., 2015) analyzed the facial features by splitting the landmarks into three groups, namely, “left eye+eyebrow”, “right eye+eyebrow” and “mouth”. Based on the statistics of the occurrence of the face regions in the CASME II database (Table 3.1), “eye+eyebrow” and “mouth” areas are the most expressive regions as they appeared the most compared to other facial regions in all the video sequences. In other words, these regions contribute the majority and meaningful micro-expression information and hence they are treated as the RoIs. Many articles in the literature demonstrated the merit of extracting the features from parts of the face rather than the whole face, as reviewed earlier in Chapter 2.2.3 and Chapter 3.3.2. The advantages and the procedure of extracting the features from the RoIs are discussed in Chapter 3.4.

The DRMF landmark detector is employed to detect 66 facial landmark coordinates, as shown in Figure 3.7(a). Then, the bounding boxes of the RoIs are determined according to the neighboring landmark points. All three RoIs (i.e., “left eye + left eyebrow”, “right eye + right eyebrow” and “mouth”) are bounded in multiple rectangular boxes, as illustrated in Figure 3.7(b). Note that, a 10 pixels margin has been added in all four directions of the boxes to encode more local expression details and to overcome the imprecise landmark annotation problem.

#### *5.5.1.2 Feature Extraction*

The LBP histograms for each RoI in each frame are calculated. Then, the apex frame is obtained by computing the correlation between the first frame (assumed to be the most

neutral expression) and the rest of the frames. The correlation coefficient is defined by:

$$d = \frac{\sum_{i=1}^{nBins} h_{1i} \times h_{2i}}{\sqrt{\sum_{i=1}^{nBins} h_{1i}^2 \times \sum_{i=1}^{nBins} h_{2i}^2}}, \quad (5.1)$$

where  $h_1$  is the gray-scale histogram of the first frame, and  $h_2$  is the current frame. Here,  $(1 - d)$  indicates the rate of difference of the LBP features between two frames. The change in differences are compared among three RoIs, and only the features of RoI with the highest change in differences is extracted for apex frame investigation.

#### 5.5.1.3 Peak Detection

Instead of spotting the apex frame by determining the maximum peak (the conventional method) (Yan, Wang, Chen, et al., 2014) in the video sequence, *divide & conquer* methodology is introduced to automatically spot the apex frame in the video sequence. Specifically, the procedures to spot the apex frame are:

1. The frame index of the peaks/ local maximum in the video sequence are detected using a peak detector.
2. The frame sequence is divided into two equal halves (e.g., a 40 frames video sequence is split into two sub-sequences containing frame 1-20 and 21-40).
3. Magnitudes of the detected peaks are summed up for each of the sub-sequence.
4. The sub-sequence with the higher magnitude will be considered for the next computation step while the other subsets will be discarded.
5. Steps (b) to (d) are repeated until the final peak (also known as apex frame) is found.

The *divide & conquer* methodology is proposed to spot the apex frame because the maximum peak of the video sequence might not necessarily represents the apex frame.

The falsely detected apex frame may be due to inaccurate feature extraction computation caused by the violation of the smoothness constraints and self-occlusions (Shreve et al., 2014). Besides, throughout a per-frame analysis in CASME II, the apex frame is likely to appear in a concentrated peaks area. Algorithm 1 shows the pseudo code of the proposed *divide & conquer* strategy.

---

**Algorithm 1** *Divide & Conquer Methodology*

---

```

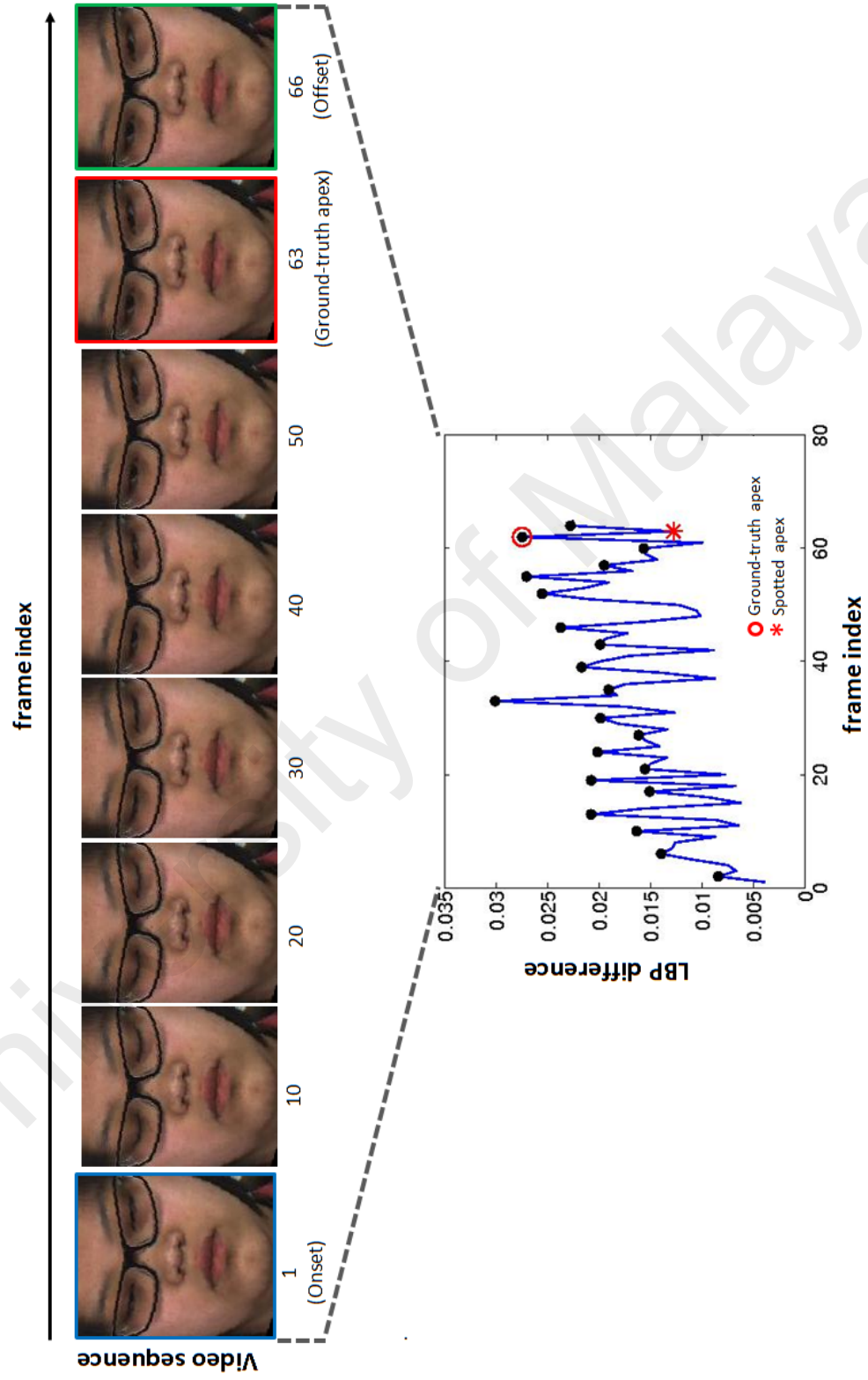
 $l \leftarrow$  split level
 $S \leftarrow$  set of candidate peaks,  $p_i$ 
Initialize  $l = 0, S_c \in \forall p_i$ 
repeat
    Split half  $S_c$  to  $S_0, S_1$ 
     $S_c \leftarrow \max(|S_0, S_1|)$ 
     $l \leftarrow l + 1$ 
until  $S_i = 1$ 

```

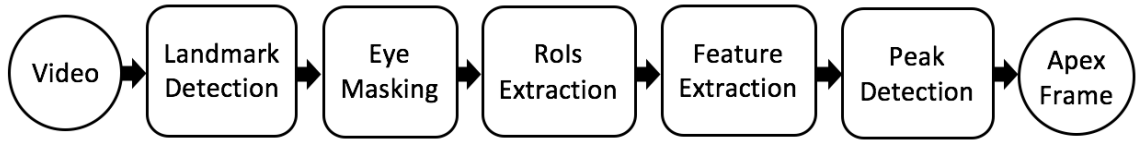
---

Figure 5.4 demonstrates the apex frame spotting approach in a sample video. It can be seen that, the ground-truth apex marked manually by the coder (frame #63) and the spotted apex by the proposed method (frame #64) differ only by one frame.





**Figure 5.4:** Demonstration of the apex frame spotting in a video sequence using LBP feature extractor with *divide & conquer* strategy



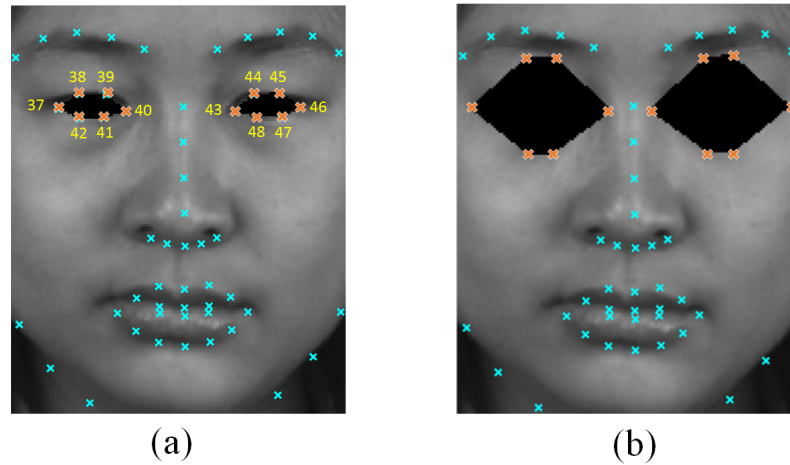
**Figure 5.5:** Flow diagram of apex frame spotting in long video

### 5.5.2 Apex Frame Spotting in Long Video

The procedure of the apex frame spotting task for long video is similar to that of short video as detailed in Chapter 5.5.1. The entire spotting procedure contains five steps, specifically:

1. Landmark detection - 66 landmark coordinates are annotated by DRMF landmark detector.
2. Eye masking - before extracting the RoIs, an eye masking approach is introduced to address the eye blinking issue.
3. RoIs extraction - three RoIs are selected, then each RoI is equally divided into multiple blocks of smaller size to encode more local appearance features.
4. Feature extraction - optical strain magnitudes are computed for each region and sum-aggregated for 12 facial blocks.
5. Peak detection - the frame with the highest sum of optical strain magnitudes (from any region) is chosen as the apex frame.

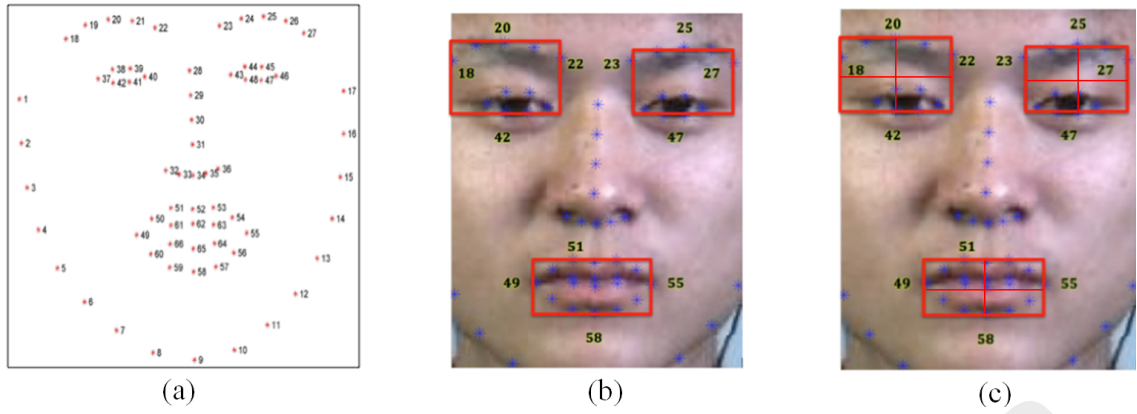
The process flow diagram of the proposed apex frame spotting approach is illustrated in Figure 5.5. Details of the aforementioned steps are elaborated in the following sub-chapters, except for step (a) landmark detection, because it is exactly the same as in Chapter 5.5.1.



**Figure 5.6:** Eye masking process: (a) There are 6 landmark coordinates which marked the boundaries of the left (landmark points 37, 38, 39, 40, 41 and 42) and 6 on the right (landmark points 43, 44, 45, 46, 47 and 48) eye regions; (b) The eye regions are removed after adding some pixel margins

#### 5.5.2.1 Eye Masking

Although eye blinking is a natural motion of rapid opening and closing of the eyelids, it is not qualified to be considered as a micro-expression. Since the micro-expression databases are typically recorded at a high frame rate, the blinking action is clearly visible when displaying the video frame-by-frame, Hence, the changes caused by eye blinking are significantly more intense compared to that of micro-expressions. Thus, it is a nagging issue that exists in some of the long video sequences. This issue is overcome by masking the left and right eye regions to reduce the false spotting of the apex frame. To ensure this is done automatically, the eye regions are removed based on the location of landmark points annotated by DRMF landmark detector (Asthana et al., 2013). The process of eliminating the eye regions is illustrated in Figure 5.6. Here, landmark coordinates 37 to 42 indicate the boundary of the right eye region, while landmark coordinates 43 to 48 are the boundary points of the left eye region. To overcome potential inaccurate landmark annotation, a 15 pixel margin is added to expand the eye boundaries.



**Figure 5.7:** Illustration of extraction of the three RoIs: (a) 66 landmark coordinates labeled by DRMF; (b) The four edges (i.e., top, bottom, left and right) are determined based on the landmark point locations; (c) Each RoI is partitioned into four blocks with the same size

#### 5.5.2.2 RoIs Extraction

The choice of the region to perform accurate spotting is crucial. Only the “eye and eye-brow” and “mouth” regions are considered for the feature extraction stage rather than the whole face region. The steps for the RoIs selection and partitioning are:

1. The facial landmark points annotated earlier (in Chapter 5.5.2.1) are adopted, where the three RoIs are identified using rectangular bounding boxes determined based on the landmark locations.
2. The RoI bounding boxes are widened by a margin of 10 pixels on all four edges to compensate for potentially imprecise landmark annotation.
3. Each RoI is equally divided into four blocks to encode more local appearance features. Thus, there is a total of 12 facial region blocks in a frame.

Figure 5.7 illustrates the steps involved in the RoIs extraction.

#### 5.5.2.3 Feature Extraction and Peak Detection

Shreve et al. (Shreve et al., 2009) employed optical strain magnitudes for macro- and micro-expression spotting. This idea is adopted to better characterize micro motions, by obtaining optical strain magnitudes based on a reference frame. TV-L1 optical flow

method (Pérez et al., 2013) is employed for the flow vector estimation. It is able to preserve flow discontinuities and is arguably more robust compared to the classical optical flow method by Black and Anandan (Black & Anandan, 1996) as employed in (Shreve et al., 2009).

The notations to be used in the subsequent chapters are first clarified. A micro-expression video clip, denoted as  $s_i$ , can be derived from Equation (3.1). Optical flow vector, namely,  $\phi$ , (from Equation (2.9)) between each frame in the video sequence (except for the first frame) and the reference frame (the first frame is chosen as it is assumed to contain the most neutral expression) is computed. Optical strain can be described by a two-dimensional displacement vector,  $\vec{u} = [u, v]$ , and its magnitude for each pixel can be calculated by taking the sum of squares of the normal and shear strain components. A more detailed discourse on optical strain can be found in Chapter 2.3.4.

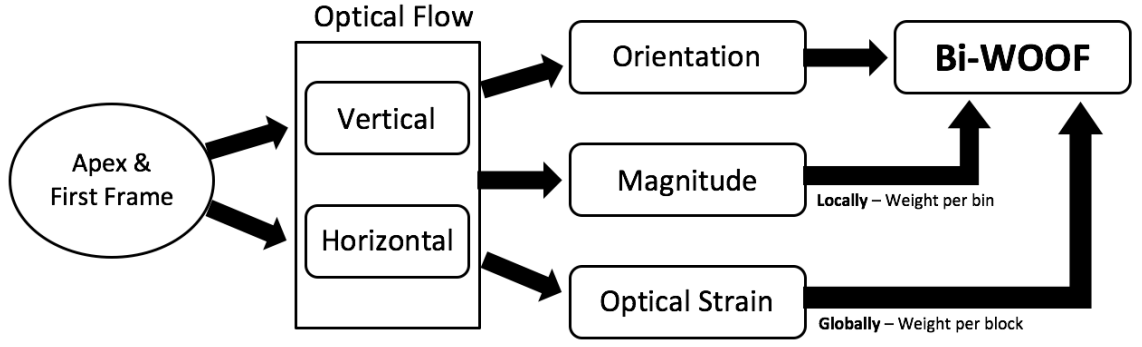
As mentioned earlier, there are 12 facial blocks in each frame. The optical strain magnitudes are calculated for each of these regions after applying eye masking. The optical strain magnitudes in each block  $b$  are summed up and the frame with the highest block value (or sum of magnitudes) is designated as the spotted apex frame  $f^*$  for the sequence:

$$f^* = \arg \max_j \left\{ \sum_{j,b} |\epsilon_{j,b}| \right\}, \quad \text{for } j \in [1, F_i - 1], b \in [1, 12]. \quad (5.2)$$

### 5.5.3 Micro-expression Recognition

A new feature descriptor, Bi-Weighted Oriented Optical Flow (Bi-WOOF) is proposed. It is capable to represent a sequence of subtle expressions by using only two frames. As illustrated in Figure 5.8, the recognition algorithm contains three main steps:

1. Estimation of optical flow - the horizontal and vertical optical flow vectors between



**Figure 5.8:** Obtaining Bi-Weighted Oriented Optical Flow (Bi-WOOF) features

the apex and first frames are estimated.

2. Computation of orientation, magnitude and optical strain - these three components are computed from the respective two optical flow components.
3. Formation of Bi-Weighted Oriented Optical Flow (Bi-WOOF) - a Bi-WOOF histogram is formed based on the orientation, with magnitude locally weighted and optical strain globally weighted.

#### 5.5.3.1 Estimation of optical flow

Optical flow approximates the changes of an object's position between two frames that are sampled at slightly different times. It encodes the motion of an object in vector notation, which indicates the direction and intensity of the flow of each image pixel.

For each video sequence,  $s_i$  (refer to Equation (3.1)), there is only one apex frame,  $f_{i,a} \in f_{i,1}, \dots, f_{i,F_i}$ , and it can be located at any frame index. The optical flow vectors of the first frame (assumed as neutral expression) and the predicted apex frames are denoted by  $f_{i,1}$  and  $f_{i,a}$ , respectively. Hence, each video of resolution  $X \times Y$  produces only one set of optical flow map, expressed as:

$$v_i = \{(u_{x,y}, v_{x,y}) | x = 1, \dots, X; y = 1, \dots, Y\}, \quad (5.3)$$

for  $i \in 1, \dots, n$ .

### 5.5.3.2 Computation of orientation, magnitude and optical strain

Given the optical flow vectors, three characteristics to describe the facial motion patterns are derived: (a) magnitude: intensity of the pixel's movement; (b) orientation: direction of the flow motion, and; (c) optical strain: subtle degree of deformation.

In order to obtain the magnitude and orientation, the flow vectors,  $\mathcal{f} = (p, q)$  (refer to Equation (2.9)), are converted from euclidean coordinates to polar coordinates:

$$\rho_{x,y} = \sqrt{p_{x,y}^2 + q_{x,y}^2}, \quad (5.4)$$

and

$$\theta_{x,y} = \tan^{-1} \frac{q_{x,y}}{p_{x,y}}, \quad (5.5)$$

where  $\rho$  and  $\theta$  are the magnitude and orientation, respectively.

The next step is to compute the optical strain,  $\varepsilon$ , based on the optical flow vectors. Refer to Chapter 2.3.4 for the derivation of optical strain magnitude.

### 5.5.3.3 Formation of Bi-Weighted Oriented Optical Flow (Bi-WOOF)

In this stage, the aforementioned characteristics (i.e., orientation, magnitude and optical strain images for every video) are utilized to build a block-based Bi-WOOF.

The three characteristic images are partitioned equally into  $N \times N$  non-overlapping blocks. For each block, the orientations  $\theta_{x,y} \in [-\pi, \pi]$  are binned and locally weighted according to its magnitude  $\rho_{x,y}$ . Thus, the range of each histogram bin is:

$$-\pi + \frac{2\pi c}{C} \leq \theta_{x,y} < -\pi + \frac{2\pi(c+1)}{C}, \quad (5.6)$$

where bin  $c \in \{1, 2, \dots, C\}$ , and  $C$  denotes the total number of histogram bins.

Next, to obtain the global weight  $\zeta_{b_1, b_2}$  for each block, the optical strain magnitude

$\epsilon_{x,y}$  are utilized as follows:

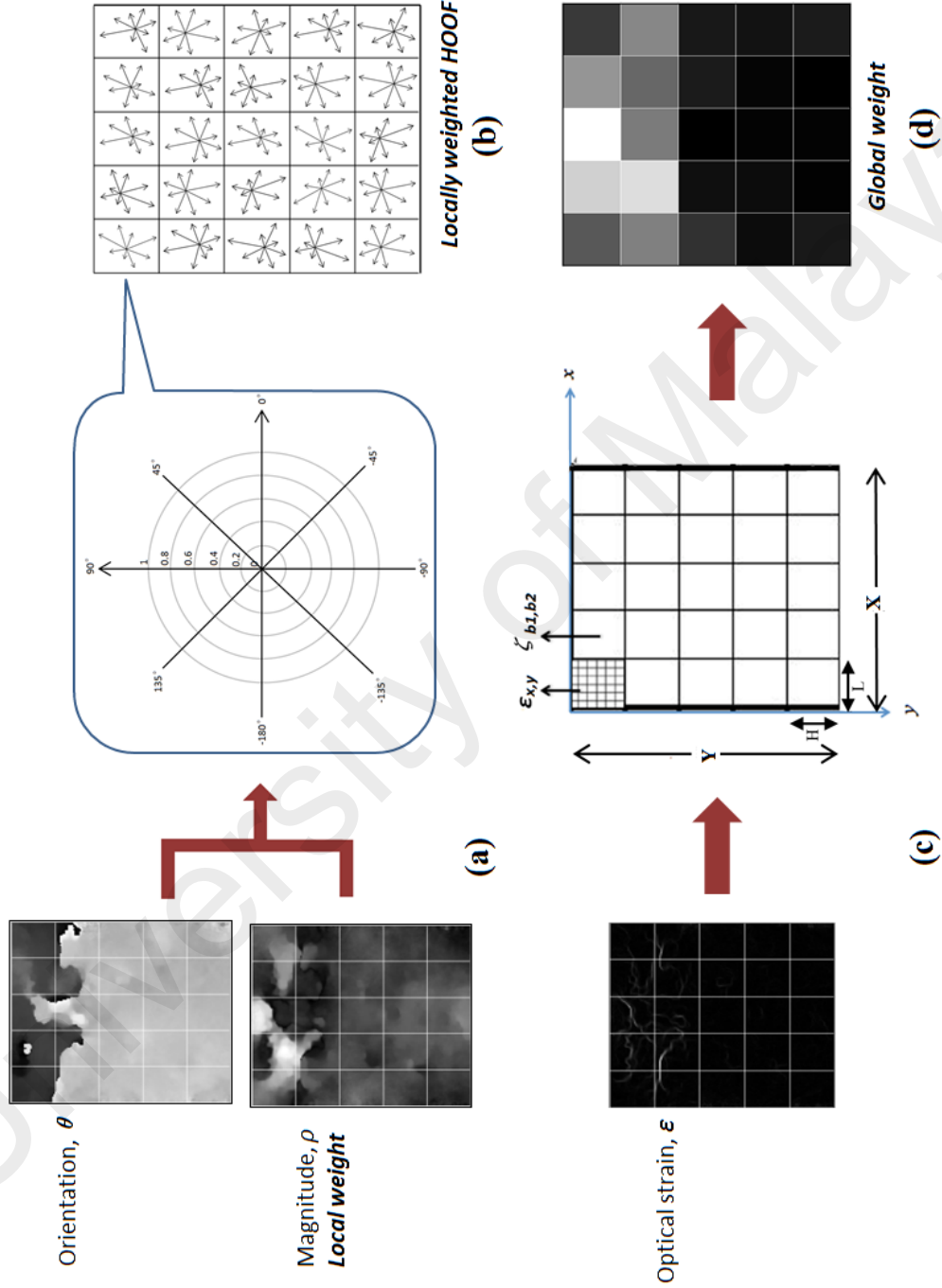
$$\zeta_{b_1,b_2} = \frac{1}{HL} \sum_{y=(b_2-1)H+1}^{b_2H} \sum_{x=(b_1-1)L+1}^{b_1L} \epsilon_{x,y}, \quad (5.7)$$

where  $L = \frac{X}{N}$ ,  $H = \frac{Y}{N}$ ,  $b_1$  and  $b_2$  are the block indices such that  $b_1, b_2 \in 1, 2, \dots, N$ , and  $X \times Y$  is the dimensions (viz., width-by-height) of the video frame. Lastly, the coefficients of  $\zeta_{b_1,b_2}$  are multiplied with the locally weighted histogram bins to their corresponding blocks. The histogram bins of each block are concatenated to form the resultant feature histogram.

Different from the conventional HOOF (Chaudhry et al., 2009) that has magnitude votes for the orientation histogram bins, both the magnitude and optical strain values are considered as the weighting schemes to highlight the importance of each optical flow. Hence, a larger intensity of the pixel's movement or deformation contributes more effect to the histogram, whereas noisy optical flows with small intensities reduce the significance of the features.

Figure 5.9 illustrates the steps in obtaining the locally and globally weighted features.





**Figure 5.9:** Bi-WOOF features formation: (a)  $\theta$  and  $\rho$  images are divided into  $N \times N$  blocks. In each block, the values of  $\rho$  for each pixel are treated as local weights to be multiplied with their respective  $\theta$  histogram bins; (b) It forms a locally weighted HOOF with feature size of  $N \times N \times C$ ; (c)  $\zeta_{b1,b2}$  denotes the global weighting matrix, which is derived from  $\epsilon$  image; (d) Finally,  $\zeta_{b1,b2}$  are multiplied with their corresponding locally weighted HOOF

## 5.6 Performance Metrics

For long videos the effectiveness of apex frame spotting can be determined using the Mean Absolute Error (MAE), which is also used in (Yan, Wang, Chen, et al., 2014). MAE indicates the average frame distance between the ground-truth and the spotted apex frame. It can be computed using the following equation:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |e_i|, \quad (5.8)$$

where  $n$  is the total number of video sequence in the database and  $e$  is the distance (in frames) between the ground-truth apex and the spotted apex. However, among the four long video databases used in the experiments, only CASME II-RAW provides ground-truth apex frame indices. Thus, to evaluate the performance of apex frame spotting, another measurement, Apex Spotting Rate (ASR) is proposed, which calculates the success rate in spotting apex frames within the onset and offset range given a long video. An apex frame is scored 1 if it is located between the onset and offset frames, and 0 otherwise:

$$\text{ASR} = \frac{1}{N} \sum_{i=1}^N \delta, \quad (5.9)$$
$$\text{where } \delta = \begin{cases} 1, & \text{if } f^* \in (f_{i,\text{onset}}, f_{i,\text{offset}}); \\ 0, & \text{otherwise.} \end{cases}$$

The classifier adopted in all experiments reported in this study is the SVM with linear kernel and LOSOCV protocol. The block size for the Bi-WOOF feature extractor is set to  $8 \times 8$  for CASME II and CASME II-RAW, while  $5 \times 5$  for the SMIC-HS, SMIC-NIR, SMIC-VIS, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR databases. The recognition accuracy is measured using the F-measure (Equation (3.5)), which conveys the balance by averaging the precision (Equation (3.4)) and recall (Equation (3.3)).

## 5.7 Results and Discussions

Two chapters are designed to discuss the experiment results for short and long video databases. Specifically, Chapter 5.7.1 reports the performance examined in the short video databases (i.e., CASME II, SMIC-HS, SMIC-NIR and SMIC-VIS), and Chapter 5.7.2 discusses the performance evaluated in the long video databases (i.e., CASME II-RAW, SMIC-E-HS, SMIC-E-NIR and SMIC-E-VIS).

### 5.7.1 Short Videos

#### 5.7.1.1 Results

The micro-expression recognition performance of the proposed method (i.e., Bi-WOOF) and the other conventional feature extraction methods are shown in Table 5.1. Note that methods #1 to #11 consider all frames in the video sequence (i.e., frames from onset to offset). However for methods #12 to #17, only two images are processed to extract the features, viz., the apex and the first frames.

To further confirm the importance of the apex frame, one frame is randomly selected in each video sequence before computing the features between that frame and the first frame using LBP, HOOF and Bi-WOOF. The recognition performances of this random frame selection approach are reported under methods #12, #14 and #16. This process is repeated for 10 times. It is observed that the utilization of apex frame always yields better recognition results when compared to the random ones. As such, it can be concluded that the apex frame plays an important role in forming discriminative features.

For method #10 (i.e., LBP-TOP), also referred to as the baseline, the experiments are re-conducted using the same datasets (4 in total) based on the original papers (Yan, Li, et al., 2014; Li et al., 2013). On the other hand, Bi-WOOF is applied to all frames in the video sequence. The features are computed by first estimating three characteristics of the optical flow (i.e., orientation, magnitude and optical strain) between the first frame and

the subsequent frames (i.e.,  $\{f_{i,1}, f_{i,j}\}, j \in 2, \dots, F_i$ ). Next, Bi-WOOF is applied for each frame in the video and in the computation of the resultant histogram. The recognition performance is reported under method #11.

For the LBP feature extractor (i.e., methods #12 and #13), the difference image is first obtained by simply performing the subtraction between the apex / random frame and the first frame. This operation aims to remove the person's identity while preserving the characteristics of facial micro-movements. Then, LBP is applied on the difference image to compute the features. Next, HOOF feature extractor (i.e., methods #14 and #15) is employed to form the histogram by binning the optical flow orientation, that is computed between the apex / random frame and the first frame. Table 5.1 suggests that the proposed algorithm (viz., method #17) achieves promising results in all four datasets. More precisely, it outperforms all the other methods in CASME II and SMIC-HS. In addition, for SMIC-VIS and SMIC-NIR, the results of the proposed method are comparable to those of method #9, viz., Xu et al. method.

**Table 5.1:** Comparison of micro-expression recognition performance in terms of F-measure on the CASME II, SMIC-HS, SMIC-VIS and SMIC-NIR databases for the state-of-the-art feature extraction methods, random frame selection approach, and the proposed algorithm

#	Methods	CASME II				SMIC-HS				SMIC-VIS				SMIC-NIR			
Whole sequence (i.e., multiple frames)	1	Le et al. (Le Ngo et al., 2014)	.33	.47	-	-	-	-	-	-	-	-	-	-	-	-	-
	2	Le et al. (Le Ngo et al., 2015)	.51	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	3	Le et al. (Le Ngo et al., 2016)	.51	.60	-	-	-	-	-	-	-	-	-	-	-	-	-
	4	Wang et al. (Y. Wang et al., 2014)	.40	.55	-	-	-	-	-	-	-	-	-	-	-	-	-
	5	Liong et al. (Liong, Phan, et al., 2014)	-	.45	-	-	-	-	-	-	-	-	-	-	-	-	-
	6	Liong et al. (Liong, See, et al., 2014)	.38	.54	-	-	-	-	-	-	-	-	-	-	-	-	-
	7	Oh et al. (Oh et al., 2015)	.43	.35	-	-	-	-	-	-	-	-	-	-	-	-	-
	8	Huang et al. (X. Huang et al., 2015)	.57	.58	-	-	-	-	-	-	-	-	-	-	-	-	-
	9	Xu et al. (Xu et al., 2016)	.30	.54	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60
	10	LBP-TOP (Yan, Li, et al., 2014; Li et al., 2013)	.39	.39	.39	.39	.39	.39	.39	.39	.39	.39	.39	.40	.40	.40	.40
	11	Bi-WOOF	.56	.53	.62	.62	.62	.62	.62	.62	.62	.62	.62	.57	.57	.57	.57
2 images (apex & first frame)	12	LBP (random & first frame)	.38	.40	.48	.48	.48	.48	.48	.48	.48	.48	.48	.51	.51	.51	.51
	13	LBP (apex & first frame)	.41	.45	.49	.49	.49	.49	.49	.49	.49	.49	.49	.54	.54	.54	.54
	14	HOOOF (random & first frame)	.41	.40	.51	.51	.51	.51	.51	.51	.51	.51	.51	.50	.50	.50	.50
	15	HOOOF (apex & first frame)	.43	.48	.49	.49	.49	.49	.49	.49	.49	.49	.49	.47	.47	.47	.47
	16	Bi-WOOF (random & first frame)	.50	.46	.56	.56	.56	.56	.56	.56	.56	.56	.56	.50	.50	.50	.50
	17	<b>Bi-WOOF (apex &amp; first frame)</b>	<b>.61</b>	<b>.62</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>	<b>.58</b>

**Table 5.2:** Confusion matrices of baseline and Bi-WOOF (apex & first frame) for recognition task on the CASME II database

(a) Baseline

	Disgust	Happiness	Others	Surprise	Repression
Disgust	<b>.20</b>	.11	.66	.02	.02
Happiness	.09	<b>.47</b>	.25	0	.19
Others	.21	.12	<b>.58</b>	.08	0
Surprise	.12	.36	.20	<b>.32</b>	0
Repression	.07	.33	.26	.04	<b>.30</b>

(b) Bi-WOOF (apex & first frame)

	Disgust	Happiness	Others	Surprise	Repression
Disgust	<b>.49</b>	.07	.44	0	0
Happiness	.03	<b>.59</b>	.28	.03	.06
Others	.21	.09	<b>.62</b>	.01	.06
Surprise	.04	.12	.08	<b>.76</b>	0
Repression	.07	.19	.22	0	<b>.52</b>

#### 5.7.1.2 Discussions

The confusion matrices for the recognition performances on the high frame rate databases, namely CASME II and SMIC-HS, are recorded in Table 5.2 and Table 5.3, respectively. It is observed that there are significant improvements in classification performance for all kinds of expression when employing Bi-WOOF (apex & first frame), compared to the baselines. In CASME II, the recognition rate of ‘Surprise’, ‘Disgust’, ‘Repression’, ‘Happiness’ and ‘Others’ expressions are improved by 44%, 29%, 22%, 12% and 4%, respectively. Furthermore, for SMIC-HS, the recognition rate of the expressions ‘Negative’, ‘Surprise’ and ‘Positive’ are improved by 32%, 19% and 18%, respectively.

**Table 5.3:** Confusion matrices of baseline and Bi-WOOF (apex & first frame) for recognition task on the SMIC-HS database

**(a)** Baseline

	Negative	Positive	Surprise
Negative	<b>.34</b>	.29	.37
Positive	.41	<b>.39</b>	.20
Surprise	.37	.19	<b>.44</b>

**(b)** Bi-WOOF (apex & first frame)

	Negative	Positive	Surprise
Negative	<b>.66</b>	.23	.11
Positive	.27	<b>.57</b>	.16
Surprise	.23	.14	<b>.63</b>



**Figure 5.10:** Illustration of components derived from optical flow using the apex and first frames of a video: (a) Horizontal vector of optical flow,  $p$ ; (b) Vertical vector of optical flow,  $q$ ; (c) Orientation,  $\theta$ ; (d) Magnitude,  $\rho$ ; (e) Optical strain,  $\epsilon$



Figure 5.10 exemplifies the components derived from optical flow using the apex frame and the first frames of the video sequence “s04\_sur\_01” in SMIC-HS, where the micro-expression of ‘Surprise’ is shown. According to the labeling criteria of emotions defined by Yan et al. (Yan, Li, et al., 2014), the changes in facial muscles are centered at the eyebrow regions. Here, the facial movements in Figure 5.10 (a), 5.10(b) and 5.10(c) are not obvious. Specifically, the whole face in the images appear to be moving and hence it is unable to determine which specific parts of the face are important. For Figure 5.10(d), a noticeable amount of muscular changes occur at the upper part of the face, whereas in Figure 5.10(e), the eyebrows regions have obvious facial movement. Since magnitude information emphasizes the amplitude of the facial changes, it is utilized as local weight. Due to higher order derivatives in obtaining the optical strain magnitudes, optical strain has the ability to remove noises and preserve large motion changes. These characteristics are manipulated to build the global weight.

Based on the results of F-measure and confusion matrices, it is observed that extracting the features of only two images (i.e., apex and first frame) using the proposed method (i.e., Bi-WOOF) is able to yield superior recognition performance for the micro-expression databases considered, especially in CASME II and SMIC-HS, which have high temporal resolution (i.e.,  $\geq 100\text{fps}$ ).

The computation time of Bi-WOOF in SMIC-HS database on both the *whole sequence* and *two images* (i.e., apex and first frame) are also examined and recorded for methods #11 and #17 in Table 5.1, respectively. The average duration taken per video for the micro-expression recognition system is 128.7134s for the *whole sequence* and 3.9499s for *two images* in MATLAB implementation. The computation time includes the time taken for execution of: (a) spotting the apex frame using the *divide & conquer* strategy; (b) estimation of the horizontal and vertical components of optical flow; (c) computation of orientation, magnitude and optical strain magnitudes; (d) generation of

Bi-WOOF histogram, and; (e) expression classification in SVM. Both experiments are carried out on an Intel Core i7-4770 CPU 3.40GHz processor. Results suggest that for the case of *two images*, it is  $\sim 33$  times faster than *whole sequence*. In other words, a speed up of  $\sim 97\%$  is achieved. It is indisputable that the method of extracting the features from only *two images* is significantly faster than using the *whole sequence*, because less images are involved in the computation and in turn, the volume of data to be processed is also less.

## 5.7.2 Long Videos

### 5.7.2.1 Results

The MAE results for apex frame spotting task on the CASME II-RAW dataset are shown in Table 5.4. It compares the techniques with and without applying eye masking using two types of feature extractors, i.e., LBP and optical strain. LBP feature is utilized in (Li et al., 2015) to spot the micro-expression frames while the use of optical strain is proposed in this study. The lower the MAE (in frames), the closer the spotted apex frame is to the ground-truth apex frame, implying more accurate spotting. It can be seen that in Table 5.4, the spotting performance of the optical strain method outperforms that of LBP. This result also emphasizes the importance of using eye masking (a more detailed look into the impact of this step can be found in Figure 5.11). Specifically, eye masking improves the spotting accuracy with optical strain features by 36.38%.

On the other hand, Table 5.5 shows the apex spotting accuracy measured in terms of ASR. With eye masking, there are tremendous improvements of 20%, 41.68%, 20.02% and 31.58% on the CASME II-RAW, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR databases, respectively. Based on these results, the elimination of eye regions (but not up to the extent of eyebrows) from consideration is able to increase the precision of searching for the apex frame. It is worth mentioning that the overall performance on the SMIC databases is

**Table 5.4:** Performance of apex frame spotting with and without eye masking on the CASME II-RAW database measured by MAE.

Feature Extractor	W/o eye mask	With eye masked	Improvement
LBP	51.86 frames	55.26 frames	-6.56%
Optical strain	42.77 frames	<b>27.21 frames</b>	<b>36.38%</b>

**Table 5.5:** Performance of apex frame spotting with and without eye masking on the long videos databases measured by ASR

Databases	W/o eye mask	With eye masked	Improvement
CASME II-RAW	0.66	<b>0.82</b>	<b>20.00%</b>
SMIC-E-HS	0.22	<b>0.38</b>	<b>41.68%</b>
SMIC-E-VIS	0.23	<b>0.28</b>	<b>20.02%</b>
SMIC-E-NIR	0.18	<b>0.27</b>	<b>31.58%</b>

still quite low (even with eye masking). This phenomenon is discussed in Chapter 5.7.2.2.

As for the performance of the recognition task, to date, there is no prior work in the literature that reports the F-measure recognition performance on micro-expression long videos (i.e., CASME II-RAW, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR). Thus, no direct comparison can be made for the F-measure results. The recognition results are tabulated in Table 5.6, where optical strain feature is employed in the spotting task. To recall, method #1 randomly spots the apex frame in the video sequence; method #2 spots the apex frame without masking the eye regions, and; method #3 spots the apex frame after applying eye masking. It can be seen that the proposed approach (method #3) achieves the best performance.

The proposed method (method #3, with eye masking) is also compared with Li et al.'s (Li et al., 2015), which is the only other work that implemented a micro-expression spotting and recognition system for long videos (see Table 5.7). However, the comparison could only be done using the Accuracy measure, and only applicable to one database, i.e., SMIC-E-VIS. It is observed that the performance of the proposed method is comparable to that of Li et al. (Li et al., 2015) but with several advantages. Specifically, the proposed method does not rely on the ground-truth onset and offset labels, and it has the

**Table 5.6:** Recognition performance for long videos in terms of F-measure

#	Methods	CASME II-RAW	SMIC-E-HS	SMIC-E-VIS	SMIC-E-NIR
1	Spotting (random) + recognition	.36	.37	.33	.28
2	Spotting (w/o eye mask) + recognition	.46	.36	.44	.38
3	<b>Spotting (with eye masked) + recognition</b>	<b>.59</b>	<b>.47</b>	<b>.53</b>	<b>.43</b>

**Table 5.7:** Comparison of the recognition accuracy between the state-of-the-art method and the proposed method on the SMIC-E-VIS database

Methods	Recognition	Remarks
Li et al. (Li et al., 2015)	.57	Onset and offset frames used. Only correctly spotted sequences used.
<b>Proposed method</b>	<b>.53</b>	No onset, offset, apex labels required. All spotted apices are used.

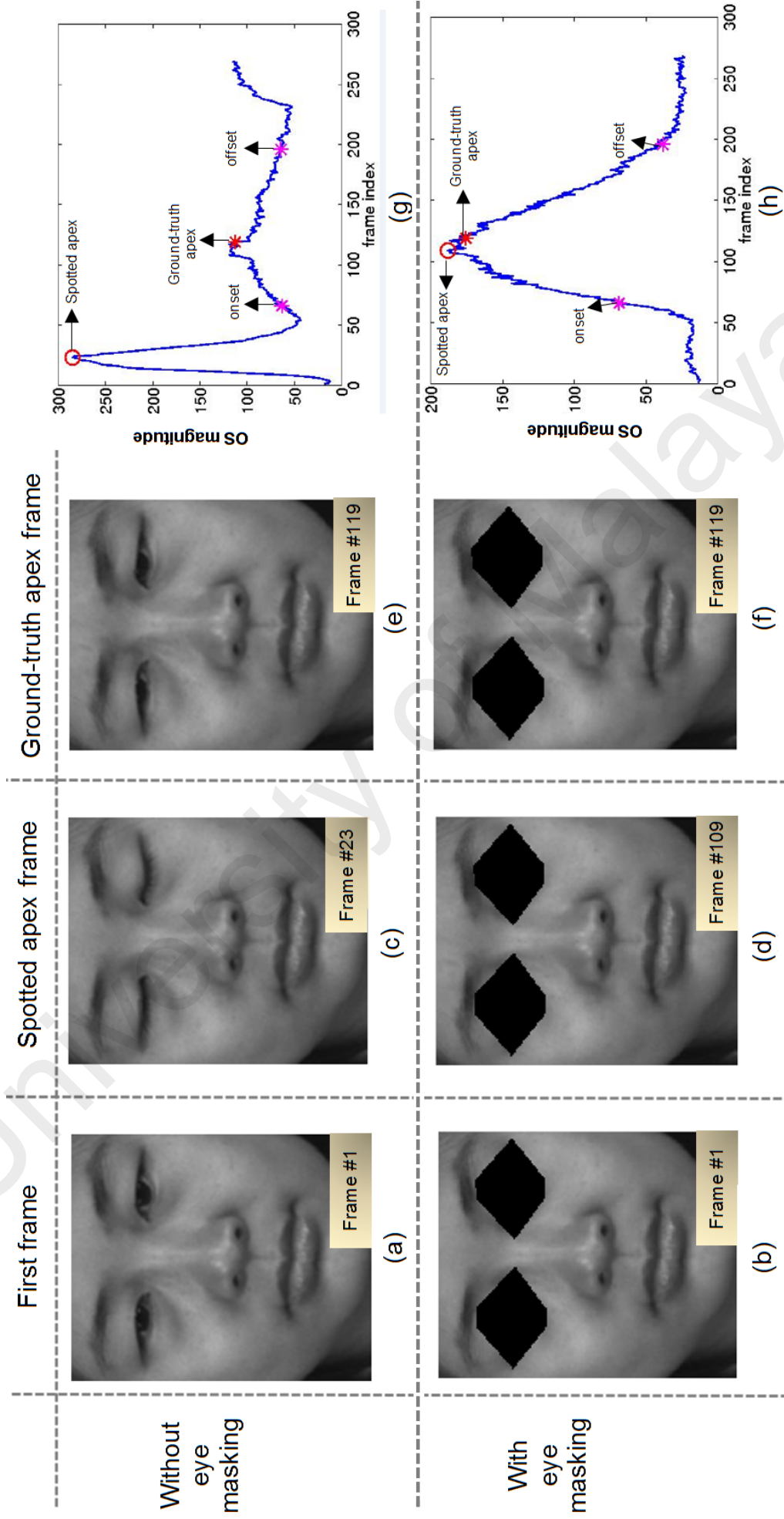
computational benefit of only needing to find the apex frame for recognition purpose.

#### 5.7.2.2 Discussions

Table 5.8 shows the average percentage of frames (among all frames in the long video) consisting of micro-expressions. Note that only approximately 6% of the frames contain micro-expressions in all three SMIC databases. In other words, 94% of the frames have either neutral faces, macro-expressions or other forms of irrelevant motions such as head rotations and eyeball movements. This suggests the possibility of macro-expressions and irrelevant movements becoming more prominent while micro-expressions may occur only in a few frames. Hence, attempting to spot the apex frame in these circumstances is an arduous task.

Nonetheless, the proposed method (method #3) is capable of coping with these issues as suggested by the results tabulated in Table 5.6 for all databases considered. For method #1, the apex frame is spotted randomly, as a control method. As expected, method #1 yields the worst recognition performance among all evaluated methods in all databases. This indicates the importance of obtaining the apex frame correctly. Both the spotting and

recognition results (in Table 5.5 and Table 5.6) support that the eye masking technique enhances the micro-expression recognition performance by removing noises from eye blinking, leading to more meaningful features. Figure 5.11 demonstrates the differences in the selection of spotted apex, with and without applying eye masking. It is observed that, without applying the eye masking technique (the upper row in Figure 5.11), the detected apex frame (frame 23) contains an eye closing motion, which is a falsely spotted micro-expression. This occurred because the facial movement is relatively more intense among all the frames in the video. On the contrary, the spotted apex frame after the application of eye masking (frame #109) is significantly closer to the ground-truth apex frame (frame #119).



**Figure 5.11:** Top row without eye masking, bottom row with eye masking: (a-b) First frame in the video; (c-d) Spotted apex frame; (e-f) Ground-truth apex frame; (g-h) Plots of optical strain magnitudes across the video sequence. Relevant frames are marked

**Table 5.8:** Average number of frames in the short and long videos of the CASME II and three SMIC databases

Databases	Short Video	Long Video	Frames with micro-expression
CASME II	67 frames	244 frames	~27%
SMIC-HS	33 frames	590 frames	~6%
SMIC-VIS	9 frames	150 frames	~6%
SMIC-NIR	9 frames	150 frames	~6%

In Table 5.7, the numerical results (spotting and recognition on the SMIC-E-VIS database) reported for Li et al.'s method are copied directly from their work (Li et al., 2015). Although their reported recognition performance is slightly better the proposed method in this study, there are several glaring differences. Firstly, they utilized the ground-truth onset and offset frame labels to form a frame interval, which is in turn used to determine the spotted micro-expression sequence. Secondly, the incorrectly spotted micro-expression sequences are not considered for recognition. The authors pointed out that the reported performance is computed by using only the correctly spotted micro-expression sequences (TPR=74.86%) (Li et al., 2015). On the other hand, the proposed approach (i.e., spotting apex frame with eye masking) eliminates the need for human intervention, or in other words, it does not make use of any hand-labeled ground-truth frames (i.e., onset, apex and offset). It also mimics a fully automatic and realistic system which considers the likelihood of a less-than-desirable spotted apex. For a closer inspection into the performance of individual classes, confusion matrices of the recognition task for CASME-II-RAW and SMIC-E-HS databases are tabulated in Table 5.9.

## 5.8 Summary

In recent years, a number of research groups attempted to improve the accuracy of micro-expression recognition by designing a variety of feature extractors that can best capture the subtle facial changes (Y. Wang et al., 2014; X. Huang et al., 2015; Y. J. Liu et al., 2016), while others (Le Ngo et al., 2015, 2016; Li et al., 2013) have sought out ways to

**Table 5.9:** Confusion matrices for the recognition task on the CASME-II-RAW and SMIC-E-HS databases using the proposed method

(a) CASME-II-RAW

	Disgust	Happiness	Tense	Surprise	Repression
Disgust	<b>.43</b>	.08	.44	.02	.03
Happiness	.06	<b>.56</b>	.25	.06	.06
Others	.23	.09	<b>.61</b>	.01	.05
Surprise	.12	.04	.16	<b>.68</b>	0
Repression	0	.07	.33	0	<b>.59</b>

(b) SMIC-E-HS

	Negative	Positive	Surprise
Negative	<b>.54</b>	.32	.14
Positive	.39	<b>.49</b>	.12
Surprise	.54	.12	<b>.34</b>

reduce information redundancy in micro-expressions (using only a portion of all frames) before recognizing them.

This chapter empirically verified that it is sufficient to encode facial micro-expression features by utilizing only the apex frame (and first frame as reference frame). Thus far, this is the first attempt at recognizing micro-expressions in video using only the apex frame. For databases that do not provide apex frame annotations, the apex frame can be acquired by automatic spotting methods. For the spotting task in long video, a major problem is the presence of eye blinking motion, which can easily be misclassified as a micro-expression. To overcome this problem, the eye regions are automatically removed by applying an automatic eye masking techniques, which depends entirely on the detected landmark coordinates. For the recognition task, a novel feature extractor is also proposed, namely Bi-Weighted Oriented Optical Flow (Bi-WOOF), which can precisely describe discriminately weighted motion features. As its name implies, the optical flow histogram features (bins) are locally weighted by their own magnitudes while facial regions (blocks) are globally weighted by the magnitude of optical strain.

Experiments are conducted on four short video micro-expression databases and four



long video databases. Among the databases tested, CASME II and SMIC-HS (short videos) achieve the highest recognition rate of 61% and 62%, respectively, when compared to the state-of-the-art methods. For the long video databases, the proposed recognition approach achieves a promising F-measure of 59% in CASME II-RAW database.

## 5.9 Prima Facie

In this work, two strong propositions are established, which are by no means conclusive at this juncture as further research is necessary:

1. **The apex frame is the most important frame in a micro-expression clip**, as it contains the most intense or expressive micro-expression information. The experiments using random frame selection (as the supposed apex frame) substantiates this fact. Perhaps, it will be interesting to know to what extent an imprecise apex frame (for example, a detected apex frame that is located a few frames away) could influence the recognition performance.
2. **The apex frame is sufficient for micro-expression recognition.** A majority of recent state-of-the-art methods promote the use of the entire video sequence, or a reduced set of frames (Li et al., 2013; Le Ngo et al., 2016). In this work, the opposite is advocated that, “less is more”, supported by the hypothesis that a large number of frames does not guarantee a high recognition accuracy, particularly in the case when high-speed cameras are employed (frame rate  $\geq 100fps$ ). Comparisons against conventional methods show that the use of a well-spotted apex frame can provide better information than an array of frames. At this juncture, it is premature to ascertain the reasons behind this finding. Hence, this warrants a detailed investigation into *how* and *where* micro-expression cues reside within the sequence itself.

## CHAPTER 6: CONCLUSION

### 6.1 Summary

A comprehensive study on micro-expression recognition system has been carried out in this dissertation. The primary focus of this research is to improve both the pre-processing and feature extraction stages in the recognition system. Experiments are performed on several latest and comprehensive spontaneous micro-expression databases. The three main contributions are highlighted as follows.

First of all, a hybrid approach to extract the important facial regions for micro-expressions recognition is proposed, namely *RoI-Selective*. It is achieved by combining both the heuristic and automatic approaches. The heuristic-based determination of salient facial regions exploits the occurrence frequency of the facial action units for all the expressions, whereas the automatic detection of the landmark points are performed using a landmark detector. The fusion of the two approaches results in the formation of the three essential RoIs (i.e., the two eye/eyebrow regions and mouth region). The *RoI-Selective* approach enhances the accuracy of the entire recognition system, by focusing the feature extraction on facial patches that contribute meaningful and expressive information. Besides, it reduces the computational complexity and hence speeds up the recognition process. Extensive experiments have also been carried out to analyze the parameter values.

Secondly, new feature descriptors are proposed by utilizing facial optical strain magnitudes to construct *Optical Strain Features (OSF)*, *Optical Strain Weighted Features (OSW)* and *Concatenation of OSF and OSW (OSF + OSW)*. Specifically, *OSF* directly utilizes the optical strain features followed by temporal sum pooling and filtering processes, whereas *OSW* adopts optical strain magnitudes as weight matrices to improve

the importance of the feature values extracted by LBP-TOP feature extractor in different block regions. The two sets of features are then concatenated to form the resultant feature histogram, i.e.,  $OSF + OSW$ . The concatenation process enriches the variety of features used, providing further robustness towards the detection and recognition of facial micro-expressions. Experiment results substantiate the capability of  $OSF + OSW$  in capturing the fine appearance and subtle muscle changes on the face.

Thirdly, a novel approach that encodes features from only two images is proposed for micro-expression recognition, in contrast to most works published in the literature that utilize either the entire video sequence or part of it for feature representation. The two images are the apex frame, which contains the highest intensity of expression changes among all frames, and the first frame of the video clip, which is assumed to have a neutral expression. To automatically spot the apex frame in the video, two approaches are designed to handle videos with different attributes, i.e., the *short videos* and *long videos*. For the *short videos*, a *divide & conquer* strategy is presented to automatically spot the apex frame. On other hand, for the *long videos*, an automated eye masking technique is proposed to exclude the eye regions in order to prevent ambiguous eye behaviors from possibly affecting the apex spotting process, and eventually the performance. After obtaining the apex frame, a new feature descriptor, called *Bi-Weighted Oriented Optical Flow (Bi-WOOF)* is introduced to encode the spotted apex frame and the first frame to represent the entire video. This method outperforms the state-of-the-art methods when considering two recent micro-expression databases.

## 6.2 Limitations

Although promising experiment results are attained, the proposed approaches in this study still pose several limitations:

1. The determination of the parameter values in the experimental settings can be te-

dious. For instance, Chapter 3 conducts a thorough analysis to empirically determine the optimal range of the parameter values, including  $r$  and  $w$  for OSF, and  $N$  and  $w$  for LBP-TOP.

2. Although the feature extractor approach introduced in Chapter 4 works well in the SMIC-HS database, there is only little improvement in CASME II on the other hand. This implies that the proposed approach might not be suitable for other micro-expression databases.
3. In Chapter 5, the first frame of the video is assumed to be the neutral frame, irrespective of short or long video sequence. This assumption can possibly be improved if the neutral frame is detected automatically prior to recognition.
4. For experiments with long video as discussed in Chapter 5, only the eye blink issues are addressed, but other motions such as: (a) neutral frame; (b) macro-expressions; (c) head movement, and; (d) other micro-expression irrelevant motions, are not handled as the currently available datasets do not contain many of these cases. Better apex spotting accuracy can potentially be achieved if large motions are dealt with.

### 6.3 Future Works

All the approaches proposed in this dissertation involve the computation of optical strain. One of the major disadvantages of utilizing optical strain is that, it is time consuming due to high complexity in deriving optical flow and optical strain values. Besides, since optical strain estimation is based upon optical flow, it is highly dependent on the brightness pattern in the image. Any subtle change in illumination, such as shadows and highlights, can lead to disastrous error in the estimation of optical strain.

On the other hand, tuning the parameters (i.e., size of the regions of interest, block partitioning of the face, etc.) towards optimum values and other settings (i.e., SVM ker-

nel) in the feature extractors as well as classifiers warrant further investigation to maximize the performance of the recognition system. Adaptive feature extractor or adaptive classifier can be introduced to automatically adjust the parameter settings in different experimental environment. In addition, better noise filtering techniques and masking of different face regions can be applied to alleviate the instability of illumination and intensity changes on the face area or background. These suggestions can potentially improve the accuracy of the computation of optical flow / strain.

Last but not least, as pointed out in Chapter 2.4, only a few of micro-expression databases are available to validate the robustness of the proposed approaches. Specifically, only the SMIC II and CASME II databases are valid for evaluation because the micro-expressions in these two databases are spontaneous, subtle, high speed, elicited in proper acquisition setup, and of sufficient number of sample size. Therefore, it is necessary to construct a new database that fulfills all the requirements for promoting and encouraging the advancement of research in micro-expression. In order to ensure practicability and relevancy, a micro-expression video should be recorded in a realistic environment with as little controlled parameters as possible.

## APPENDIX A: LIST OF PUBLICATIONS AND PAPERS PRESENTED

The following is the list of submitted / accepted journal articles and peer-viewed conference papers related to this study.

### **Journals:**

[1] Liong, Sze-Teng, John See, Raphael Chung-Wei Phan, Yee-Hui Oh, Anh Cat Le Ngo, KokSheik Wong, and Su-Wei Tan. (2016). Spontaneous Subtle Expression Detection and Recognition based on Facial Strain. *Signal Processing: Image Communication*, Volume 47, (pp. 170–182). Doi: <http://dx.doi.org/10.1016/j.image.2016.06.004> (impact factor 2015: 1.602).

[2] Liong, Sze-Teng, John See, Raphael Chung-Wei Phan, and KokSheik Wong. (2016). Less is More: Micro-expression Recognition from Video using Apex Frame. *Neurocomputing*. (Submitted).

[3] Liong, Sze-Teng, John See, Raphael Chung-Wei Phan, KokSheik Wong. Tan, Su-Wei. (2016). Hybrid Facial Regions Extractions for Micro-expression Recognition System. *Journal Signal Processing Systems*. (Submitted on 2 February 2016, revised on 20 October 2016).

### **International Peer-Reviewed Conferences:**

[1] Liong, Sze-Teng, John See, Raphael C-W. Phan, Anh Cat Le Ngo, Yee-Hui Oh, and KokSheik Wong. (2014). Subtle expression recognition using optical strain weighted features. In *Asian Conference on Computer Vision*, (pp. 644–657). Doi: [http://dx.doi.org/10.1007/978-3-319-16631-5\\_47](http://dx.doi.org/10.1007/978-3-319-16631-5_47)

[2] Liong, Sze-Teng, Raphael C-W. Phan, John See, Yee-Hui Oh, and KokSheik Wong. (2014). Optical strain based recognition of subtle emotions. In *Intelligent Signal Processing and Communication Systems*, (pp. 180–184). Doi: <http://dx.doi.org/10.1109/ISPACS.2014.70244>

[3] Oh, Yee-Hui, Anh Cat Le Ngo, John See, Sze-Teng Liong, Raphael C-W. Phan, and

Huo-Chong Ling. (2015). Monogenic riesz wavelet representation for micro-expression recognition. In IEEE International Conference on Digital Signal Processing, (pp. 1237–1241). Doi: <http://dx.doi.org/10.1109/10.1109/ICDSP.2015.7252078>

[4] Le Ngo, Anh Cat, Sze-Teng Liong, John See, and Raphael Chung-Wei Phan. (2015). Are subtle expressions too sparse to recognize? In IEEE International Conference on Digital Signal Processing, (pp. 1246–1250). Doi: <http://dx.doi.org/10.1109/ICDSP.2015.7252080>

[5] Liong, Sze-Teng, John See, KokSheik Wong, Anh Cat Le Ngo, Yee-Hui Oh, and Raphael Phan. (2015). Automatic Apex Frame Spotting in Micro-expression Database. In 3rd IAPR Asian Conference on Pattern Recognition, (pp. 665 - 669). Doi: <http://dx.doi.org/10.1109>

[6] Liong, Sze-Teng, John See, KokSheik Wong and Raphael Chung-Wei Phan. (2016). Automatic Micro-expression Recognition from Long Video using a Single Spotted Apex. In Asian Conference on Computer Vision Workshop. (Accepted).

## REFERENCES

- Ahonen, T., Hadid, A., & Pietikäinen, M. (2004, May). Face recognition with local binary patterns. In *European Conference on Computer Vision* (pp. 469–481).
- Anandan, P. (1987). Measuring visual motion from image sequences..
- Anderson, K., & McOwan, P. W. (2006). A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(1), 96–105.
- Asthana, A. (n.d.). *Discriminative response map fitting (drmf 2013)*. <http://ibug.doc.ic.ac.uk/resources>. (Accessed: 2013-8-27)
- Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3444–3451).
- Barcelos, C. A. Z., Boaventura, M., & Silva Jr, E. C. (2003). A well-balanced flow equation for noise removal and edge detection. *IEEE Transactions on Image Processing*, 12(7), 751–763.
- Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1), 43–77.
- Beauchemin, S. S., & Barron, J. L. (1995). The computation of optical flow. *ACM Computing Surveys*, 27(3), 433–466.
- Black, M. J., & Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1), 75–104.
- Boureau, Y. L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *International Conference on Machine Learning* (pp. 111–118).
- Brahnam, S., Nanni, L., & Sexton, R. (2007). Introduction to neonatal facial pain detection using common and advanced face classification techniques. In *Advanced Computational Intelligence Paradigms in Healthcare — 1* (pp. 225–253). Springer Berlin Heidelberg.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Chatterjee, P., & Milanfar, P. (2012). Patch-based near-optimal image denoising. *IEEE Transactions on Image Processing*, 21(4), 1635–1649.
- Chaudhry, R., Ravichandran, A., Hager, G., & Vidal, R. (2009, June). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision and*



- Chebira, A., Barbotin, Y., Jackson, C., Merryman, T., Srinivasa, G., Murphy, R. F., & Kovačević, J. (2007). A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, 8(1), 1.
- Chingovska, I., Anjos, A., & Marcel, S. (2013). Anti-spoofing in action: joint operation with a verification system. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 98–104).
- Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998). Active appearance models. In *European Conference on Computer Vision* (pp. 484–498).
- Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1), 38–59.
- Cristinacce, D., & Cootes, T. (2008). Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10), 3054–3067.
- Cristinacce, D., & Cootes, T. F. (2004, May). A comparison of shape constrained facial feature detectors. In *Automatic face and gesture recognition* (Vol. 2(5), pp. 375–380).
- Cristinacce, D., & Cootes, T. F. (2006, September). Feature detection and tracking with constrained local models. In *British Machine Vision Conference* (Vol. 2(5), p. 6).
- Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 886–893).
- Davison, A. K., Yap, M. H., Costen, N., Tan, K., Lansley, C., & Leightley, D. (2014, September). Micro-facial movements: An investigation on spatio-temporal descriptors. In *European Conference on Computer Vision* (pp. 111–123).
- Davison, A. K., Yap, M. H., & Lansley, C. (2015, October). Micro-facial movement detection using individualised baselines and histogram-based descriptors. In *IEEE International Conference on Systems, Man, and Cybernetics* (pp. 1864–1869).
- Dawood, M., Gigengack, F., Jiang, X., & Schäfers, K. P. (2013). A mass conservation-based optical flow method for cardiac motion correction in 3D-PET. *Medical Physics*, 40(1), 217–226.
- Ekman, P. (2009a). Lie catching and microexpressions. In (pp. 118–133). Oxford University.
- Ekman, P. (2009b). Telling lies: Clues to deceit in the marketplace, politics, and marriage. WW Norton and Company.
- Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, 32, 88–106.

- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124.
- Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology*, 29(3), 288.
- Ekman, P., & Friesen, W. V. (1978). *Facial action coding system*. Consulting Psychologists Press.
- Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system*. Salt Lake City: Research Nexus eBook.
- Fan, X., & Verma, B. (2009). Selection and fusion of facial features for face recognition. *Expert systems with applications*, 36(3), 7157–7169.
- Fleet, D., & Weiss, Y. (2006). Optical flow estimation. *Handbook of Mathematical Models in Computer Vision*, 237–257.
- Fleet, D. J., & Jepson, A. D. (1990). Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1), 77–104.
- Forsyth, D. A., & Ponce, J. (2002). *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference.
- Frank, M. G., Herbasz, M., Sinuk, K., Keller, A., Kurylo, A., & Nolan, C. (2009). I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In *The Annual meeting of the International Communication Association, Sheraton New York, New York City, NY*.
- Frank, M. G., Maccario, C. J., & Govindaraju, V. (2009). Protecting airline passengers in the age of terrorism. In (p. 86—106). ABC-CLIO.
- Gao, W., Zhang, X., Yang, L., & Liu, H. (2010). An improved sobel edge detection. In *IEEE International Conference on Computer Science and Information Technology* (Vol. 5, pp. 67–71).
- Garcia, F., Cerri, P., Broggi, A., de la Escalera, A., & Armingol, J. M. (2012, June). Data fusion for overtaking vehicle detection based on radar and optical flow. In *IEEE Intelligent Vehicles Symposium (IV)* (pp. 494–499).
- Gatos, B., Pratikakis, I., & Perantonis, S. J. (2004, September). An adaptive binarization technique for low quality historical documents. In *International Workshop on Document Analysis Systems* (pp. 102–113).
- Gibson, J. J. (1950). *The perception of the visual world*. Greenwood.
- Goldstein, J. S., Reed, I. S., & Scharf, L. L. (1998). A multistage representation of the Wiener filter based on orthogonal projections. *IEEE Transactions on Information Theory*, 44(7), 2943–2959.
- Goshtasby, A. (1988). Image registration by local approximation methods. *Image and*

- Guo, Z., Zhang, L., & Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6), 1657–1663.
- Haggard, E. A., & Isaacs, K. S. (1966). Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. *Methods of Research in Psychotherapy*, 154–165.
- Hamel, P., Lemieux, S., Bengio, Y., & Eck, D. (2011, October). Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In *International Society for Music Information Retrieval Conference* (pp. 729–734).
- Hamilton, O. K., & Breckon, T. P. (2016, September). Generalized dynamic object removal for dense stereo vision based scene mapping using synthesised optical flow. In *IEEE International Conference on Digital Image Processing* (pp. 3439–3443).
- Happy, S. L., & Routray, A. (2015). Automatic facial expression recognition using features of salient facial patches. *IEEE Transaction on Affective Computing*, 6(1), 1–12.
- Hayat, M., Bennamoun, M., & El-Sallam, A. (2012, June). Evaluation of spatiotemporal detectors and descriptors for facial expression recognition. In *IEEE International Conference on Human System Interactions* (pp. 43–47).
- Heeger, D. J. (1987). Model for the extraction of image flow. *Journal of the Optical Society of America A*, 4(8), 1455–1471.
- Heimdal, A., Støylen, A., Torp, H., & Skjærpe, T. (1998). Real-time strain rate imaging of the left ventricle by ultrasound. *Journal of the American Society of Echocardiography*, 11(11), 1013–1019.
- Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. In *Artificial intelligence* (Vol. 17, pp. 185–203).
- Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2004, July). Extreme learning machine: a new learning scheme of feedforward neural networks. In *IEEE International Joint Conference on Neural Networks* (Vol. 2, pp. 985–990).
- Huang, X., Wang, S. J., Zhao, G., & Piteikainen, M. (2015, September). Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In *IEEE International Conference on Computer Vision Workshops* (pp. 1–9).
- Jantzen, J., Norup, J., Dounias, G., & Bjerregaard, B. (2005). Pap-smear benchmark data for pattern classification. *Nature inspired Smart Information Systems*, 8(1), 1–9.
- Jiang, B., Martinez, B., Valstar, M. F., & Pantic, M. (2014, August). Decision level fusion of domain specific regions for facial action recognition. In *International Conference on Pattern Recognition*.

- Jin, Y., Cao, J., Ruan, Q., & Wang, X. (2014). Cross-modality 2D-3D face recognition via multiview smooth discriminant analysis based on ELM. *Journal of Electrical and Computer Engineering*, 21.
- Joshi, J., Dhall, A., Goecke, R., Breakspear, M., & Parker, G. (2012, November). Neural-net classification for spatio-temporal descriptor based depression analysis. In *IEEE International Conference on Pattern Recognition* (pp. 2634–2638).
- Juneja, M., & Sandhu, P. S. (2009). Performance evaluation of edge detection techniques for images in spatial domain. *International Journal of Computer Theory and Engineering*, 1(5), 614.
- Kellokumpu, V., Zhao, G., Li, S. Z., & Pietikäinen, M. (2009). Dynamic texture based gait recognition. In *International Conference on Biometrics* (pp. 1000–1009).
- Kenji, M. A. S. E. (1991). Recognition of facial expression from optical flow. *IEICE Transactions on Information and Systems*, 74(10), 3474–3483.
- Klaser, A., Marszałek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3D-gradients. In *British machine vision conference* (p. 275-1).
- Kovashka, A., & Grauman, K. (2010, June). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2046–2053).
- Lamsal, B., & Matsumoto, N. (2015). Effects of the Unscented Kalman filter process for high performance face detector. *International Journal of Information and Electronics Engineering*, 5(6), 454.
- Lee, E. H. (1969). Elastic-plastic deformation at finite strains. *Journal of Applied Mechanics*, 36(1), 1–6.
- Le Ngo, A. C., Liong, S. T., See, J., & Phan, R. C. W. (2015, July). Are subtle expressions too sparse to recognize? In *IEEE International Conference on Digital Signal Processing* (pp. 1246–1250).
- Le Ngo, A. C., Phan, R. C. W., & See, J. (2014, November). Spontaneous subtle expression recognition: Imbalanced databases and solutions. In *Asian Conference on Computer Vision* (pp. 33–48).
- Le Ngo, A. C., See, J., & Phan, R. C. W. (2016). Sparsity in dynamics of spontaneous subtle emotions: Analysis & application. *IEEE Transactions on Affective Computing*.
- Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., & Pietikäinen, M. (2015). Reading hidden emotions: Spontaneous micro-expression spotting and recognition. *arXiv preprint arXiv:1511.00423*.
- Li, X., Pfister, T., Huang, X., Zhao, G., & Pietikainen, M. (2013, April). A spontaneous micro-expression database: Inducement, collection and baseline. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (pp.

1–6).

- Lien, J. J. J., Kanade, T., Cohn, J. F., & Li, C. C. (2000). Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3), 131–146.
- Liong, S. T., Phan, R. C. W., See, J., Oh, Y. H., & Wong, K. (2014). Optical strain based recognition of subtle emotions. In *International Symposium on International Symposium on Intelligent Signal Processing and Communication Systems* (pp. 180–184).
- Liong, S. T., See, J., Phan, R. C. W., Le Ngo, A. C., Oh, Y. H., & Wong, K. (2014). Subtle expression recognition using optical strain weighted features. In *Asian Conference on Computer Vision*.
- Liu, Y. J., Zhang, J. K., Yan, W. J., Wang, S. J., Zhao, G., & Fu, X. (2016). A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing*, To appear.
- Liu, Z., Shan, Y., & Zhang, Z. (2001, August). Expressive expression mapping with ratio images. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 271–276).
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110.
- Lucas, B. D., & Kanade, T. (1981, August). An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence* (Vol. 81(1), pp. 674–679).
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops* (pp. 94–101).
- Lyons, M. J., Akamatsu, S., Kamachi, M., Gyoba, J., & Budynek, J. (1998). The Japanese female facial expression (JAFFE) database.
- Lyons, M. J., Budynek, J., & Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12), 1357–1362.
- Ma, Y., & Cisar, P. (2009, June). Event detection using local binary pattern based dynamic textures. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 38–44).
- Manohar, V., Goldgof, D., Sarkar, S., & Zhang, Y. (2007, February). Facial strain pattern as a soft forensic evidence. In *IEEE Workshop on Applications of Computer Vision* (pp. 42–42).
- Mattivi, R., & Shao, L. (2009, September). Human action recognition using LBP-TOP as

- sparse spatio-temporal feature descriptor. In *International Conference on Computer Analysis of Images and Patterns* (pp. 740–747).
- Mirmohamadsadeghi, L., & Drygajlo, A. (2011, October). Palm vein recognition with local binary patterns and local derivative patterns. In *IEEE International Joint Conference on Biometrics* (pp. 1–6).
- Moilanen, A., Zhao, G., & Pietikainen, M. (2014, August). Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *IEEE International Conference on Pattern Recognition* (pp. 1722–1727).
- Mu, Y., Yan, S., Liu, Y., Huang, T., & Zhou, B. (2008, June). Discriminative local binary patterns for human detection in personal album. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8).
- Nagel, H. H. (1983). Displacement vectors derived from second-order intensity variations in image sequences. *Computer Vision, Graphics, and Image Processing*, 21(1), 85–117.
- Nanni, L., Lumini, A., & Brahnam, S. (2010). Local binary patterns variants as texture descriptors for medical image analysis. *Artificial Intelligence in Medicine*, 49(2), 117–125.
- Niu, Z., Shan, S., Yan, S., Chen, X., & Gao, W. (2006). 2d cascaded adaboost for eye localization. In *IEEE International Conference on Pattern Recognition* (Vol. 2, pp. 1216–1219).
- Oh, Y. H., Le Ngo, A. C., See, J., Liong, S. T., Phan, R. C. W., & Ling, H. C. (2015, July). Monogenic riesz wavelet representation for micro-expression recognition. In *IEEE International Conference on Digital Signal Processing* (pp. 1237–1241).
- Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 971–987.
- Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1), 51–59.
- O’Sullivan, M., Frank, M. G., Hurley, C. M., & Tiwana, J. (2009). Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6), 530–538.
- Pai, N. S., & Chang, S. P. (2011). An embedded system for real-time facial expression recognition based on the extension theory. *Computers & Mathematics with Applications*, 61(8), 2101–2106.
- Pfister, T., Li, X., Zhao, G., & Pietikainen, M. (2011, November). Recognising spontaneous facial micro-expressions. In *IEEE International Conference In Computer Vision* (pp. 1449–1456).
- Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and

- evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5), 295–306.
- Polikovskiy, S., Kameda, Y., & Ohta, Y. (2009, December). Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In *International Conference on Crime Detection and Prevention* (pp. 1–6).
- Porter, S., & Ten Brinke, L. (2008). Reading between the lies identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, 19(5), 508–514.
- Pérez, J. S., Meinhardt-Llopis, E., & Facciolo, G. (2013). TV-L1 optical flow estimation. *Image Processing On Line*, 137–150.
- Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalande, & Pantic, M. (2015). The AV+EC 2015 multimodal affect recognition challenge: Bridging across audio, video, and physiological data. In *ACM International Workshop on Audio/Visual Emotion Challenge*.
- Rudovic, O., Pavlovic, V., & Pantic, M. (2014). Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5), 944–958.
- Samaria, F. S., & Harter, A. C. (1994, December). Parameterisation of a stochastic model for human face identification. In *IEEE Workshop on Applications of Computer Vision* (pp. 138–142).
- Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2), 200–215.
- Savran, A., Sankur, B., & Bilge, M. T. (2012). Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10), 774–784.
- Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3), 297–336.
- Schwarz, L. A., Mkhitarian, A., Mateus, D., & Navab, N. (2012). Human skeleton tracking from depth data using geodesic distances and optical flow. *Image and Vision Computing*, 30(3), 217–226.
- Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6), 803–816.
- Shan, C., & Gritti, T. (2008, September). Learning discriminative LBP-histogram bins for facial expression recognition. In *British Machine Vision Conference*.
- Shreve, M., Brizzi, J., Fefilatyev, S., Luguev, T., Goldgof, D., & Sarkar, S. (2014). Automatic expression spotting in videos. *Image and Vision Computing*, 32(8), 476–486.

- Shreve, M., Godavarthy, S., Goldgof, D., & Sarkar, S. (2011, March). Macro-and micro-expression spotting in long videos using spatio-temporal strain. In *IEEE International Conference on Automatic Face, Gesture Recognition and Workshops* (pp. 51–56).
- Shreve, M., Godavarthy, S., Manohar, V., Goldgof, D., & Sarkar, S. (2009). Towards macro-and micro-expression spotting in video using strain patterns. In *IEEE Workshop on Applications of Computer Vision* (pp. 1–6).
- Shreve, M., Jain, N., Goldgof, D., Sarkar, S., Kropatsch, W., Tzou, C. H. J., & Frey, M. (2011, January). Evaluation of facial reconstructive surgery on patients with facial palsy using optical strain. In *Computer Analysis of Images and Patterns* (pp. 512–519).
- Shreve, M., Manohar, V., Goldgof, D., & Sarkar, S. (2010). Face recognition under camouflage and adverse illumination. In *IEEE International Conference on Biometrics: Theory Applications and Systems* (pp. 1–6).
- Simof, J. C., & Hughes, T. J. (2008). *Computational inelasticity*. Springer.
- Singh, A. (1990, December). An estimation-theoretic framework for image-flow computation. In *IEEE International Conference on Computer Vision* (pp. 168–177).
- Sivic, J., & Zisserman, A. (2003, October). Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision* (pp. 1470–1477).
- Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3), 293–300.
- Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook* (pp. 667–685). Springer US.
- Valstar, M., & Pantic, M. (2006, June). Fully automatic facial action unit detection and temporal analysis. In *Conference on Computer Vision and Pattern Recognition Workshop*.
- Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., . . . Cohn, J. F. (2015, may). Fera 2015-second facial expression recognition and analysis challenge. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (Vol. 6, pp. 1–8).
- Valstar, M. F., & Pantic, M. (2007). Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *International Workshop on Human-Computer Interaction* (pp. 118–127).
- Van Ginneken, B., Frangi, A. F., Staal, J. J., ter Haar Romeny, B. M., & Viergever, M. A. (2002). Active shape model segmentation with optimal features. *IEEE Transaction on Medical Imaging*, 21(8), 924–933.
- Varma, M., & Ray, D. (2007, October). Learning the discriminative power-invariance



- trade-off. In *IEEE International Conference on Computer Vision* (pp. 1–8).
- Vinciarelli, A., Dielmann, A., Favre, S., & Salamin, H. (2009, September). Canal9: A database of political debates for analysis of social interactions. In *International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1–4).
- Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137–154.
- Wang, H., Ullah, M. M., Klaser, A., Laptev, I., & Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *British machine vision conference* (p. 124-1).
- Wang, S., Yan, W., Li, X., Zhao, G., Zhou, C., Fu, X., ... Tao, J. (2015). Micro-expression recognition using color spaces. *IEEE Transactions on Image Processing*, 24(12), 6034–6047.
- Wang, S. J., Chen, H. L., Yan, W. J., Chen, Y. H., & Fu, X. (2014). Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural Processing Letters*, 39(1), 25–43.
- Wang, S. J., Yan, W. J., Zhao, G., Fu, X., & Zhou, C. G. (2014). Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In *Workshop at the European Conference on Computer Vision* (pp. 325–338).
- Wang, S. J., Zhou, C. G., Zhang, X. J., N.and Peng, Chen, Y. H., & Liu, X. (2011). Face recognition using second-order discriminant tensor subspace analysis. *Neurocomputing*, 74(12), 2142–2156.
- Wang, Y., See, J., Phan, R. C. W., & Oh, Y. H. (2014, November). LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition. In *Asian Conference on Computer Vision* (pp. 525–537).
- Warren, G., Schertler, E., & Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33(1), 59–69.
- Xi, M., Chen, L., Polajnar, D., & Tong, W. (2016, September). Local binary pattern network: A deep learning approach for face recognition. In *IEEE International Conference on Digital Image Processing* (pp. 3224–3228).
- Xu, F., Zhang, J., & Wang, J. (2016). Microexpression identification and categorization using a facial dynamics map. *IEEE Transactions on Affective Computing*.
- Yamaji, A. (2007). *An introduction to tectonophysics: Theoretical aspects of structural geology*. TERRAPUB.
- Yan, W. J., Li, X., Wang, S. J., Zhao, G., Liu, Y. J., H., C. Y., & Fu, X. (2014). CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1), e86041.

- Yan, W. J., Wang, S. J., Chen, Y. H., Zhao, G., & Fu, X. (2014, September). Quantifying micro-expressions with constraint local model and local binary pattern. In *Workshop at the European Conference on Computer Vision* (pp. 296–305).
- Yan, W. J., Wang, S. J., Liu, Y. J., Wu, Q., & Fu, X. (2014). For micro-expression recognition: Database and suggestions. *Neurocomputing*, 136, 82–87.
- Yan, W. J., Wu, Q., Liu, Y. J., Wang, S. J., & Fu, X. (2013). CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (pp. 1–7).
- Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2), 213–238.
- Zhang, Y., & Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5), 699–714.
- Zhao, G., Barnard, M., & Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transaction on Multimedia*, 11(7), 1254–1265.
- Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 29(6), 915–928.
- Zhao, G., & Pietikäinen, M. (2009). Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. *Pattern Recognition Letters*, 30(12), 1117–1127.
- Zhao, G., & Pietikäinen, M. (2013, June). Visual speaker identification with spatiotemporal directional features. In *International Conference Image Analysis and Recognition* (pp. 1–10).
- Zhong, L., Liu, Y., Q., P., J., Huang, & Metaxas, D. N. (2015). Learning multiscale active facial patches for expression analysis. *IEEE Transactions on Cybernetics*, 45(8), 1499–510.
- Zhou, Z., Zhao, G., Guo, Y., & Pietikainen, M. (2012). An image-based visual speech animation system. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(10), 1420–1432.
- Zhu, X., & Ramanan, D. (2012, June). Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2879–2886).